The Bayley-III-NL special needs addition

Visser, Linda

Publication date:
2014

The Bayley-III-NL Special Needs Addition

A suitable developmental assessment instrument for young children with special needs

Linda Visser

The work presented in this dissertation was supported by a grant from ZonMw – the Netherlands Organization for Health Research and Development (grant 157013002).

This thesis has been supported by the graduate school of Behavioural and Cognitive Neurosciences in Groningen.

ISBN: 978-90-5963-072-7
ISBN: 978-90-5963-073-4 (electronic version)
NUR: 848

Stichting Kinderstudies
Grote Kruisstraat 2/1
9712 TJ Groningen
The Netherlands
e-mail: b.f.van.der.meulen@rug.nl
website: www.stichting-kinderstudies.nl

Cover design: Maaike van den Heuvel & Gerco Hiddink, Hartebeest
Printed by: Wöhrmann Print Service

The Bayley-III-NL Special Needs Addition

A suitable developmental assessment instrument for young children with special needs

Proefschrift

ter verkrijging van de graad van doctor aan de
Rijksuniversiteit Groningen
op gezag van de
rector magnificus, prof. dr. E. Sterken
en volgens besluit van het College voor Promoties.

De openbare verdediging zal plaatsvinden op

donderdag 9 januari 2014 om 14:30 uur

door

Linda Visser

geboren op 21 mei 1985
te Woerden

Promotor:
Prof. dr. A.J.J.M. Ruijssenaars

Copromotores:
Dr. M.E. Timmerman
Dr. S.A.J. Ruiter

Beoordelingscommissie:
Prof. dr. A.L. van Baar
Prof. dr. R.R. Meijer
Prof. dr. H.J.M. Janssen

Contents

Chapter 1

Introduction

1.1 Introduction

Incidence estimates of special needs in children range from 5 to 10% in western countries (Rydz, Shevell, Majnemer, & Oskoui, 2005). Special needs are defined as a physical or mental impairment that substantially limits one or more of the major life activities of an individual (Americans with Disabilities Act, 1990). This includes general developmental delay, disorders such as pervasive developmental disorder and Down syndrome, and functional disabilities such as motor and visual impairments.

In cases of special needs, it is important to intervene early in life, because brain plasticity is highest in the first few years of life (Rimrodt & Johnston, 2009). This means that the development of a very young child is highly susceptible to environmental influences, while this susceptibility decreases with age. Even though the unpredictability of early development makes it difficult to study the effects of intervention as a function of age, there is consensus that early intervention is more effective than later intervention (Shonkoff & Phillips, 2000).

As part of early intervention, developmental assessment generally takes place in order to estimate and track the level of development in different areas and thus optimally tailor the intervention to the abilities of the child. The contemporary view of development is that it is the result of continuous transactions between the child and different parts of his/her environment (see, for example, Sameroff, 2009). In accordance with this view, it is often argued that developmental information should be repeatedly collected, and from different sources, by combining different assessment methods, such as observation, parent interview, and standardized developmental assessment (Individuals with Disabilities Education Improvement Act, 2004a).

The current thesis focuses on standardized developmental assessment. A standardized developmental assessment instrument enables a child's developmental level to be estimated. This is done by administering standardized test tasks to a child and comparing the resulting test score to the test score distribution of a norm group. The norm group usually represents the population of typically developing children of the same age. One widely used instrument is the Bayley Scales of Infant and Toddler Development, Third Edition (Bayley-III; Bayley, 2006a). It assesses the cognitive, language, and motor development of children between one and 42 months of age. The instrument is eclectic: It has historically been developed from a variety of different scales of infant development and infant and toddler research, as described in the technical manual (Bayley, 2006b). In this newest version older concepts, such as pretend play and novelty preference, are combined with concepts that are based on more recent studies on, for example, information processing (Albers & Grieve, 2007). The Bayley-III has been adapted for the Dutch target group, resulting in the Bayley-III-NL (Steenis, Verhoeven, & Van Baar, 2012). Standardization research of the Bayley-III-NL is currently ongoing and will be finished in 2014.

Standardized developmental assessment instruments like the Bayley-III thus play an important role in early intervention. However, in some cases problems arise with respect to the validity and usefulness of the results of standardized developmental assessment (Hebbeler, Barton, & Mallik, 2008; Neisworth & Bagnato, 2004). The current thesis focuses on two of these problems which are related to the influence of functional impairments on test results (Section 1.2) and the use of developmental assessment results as a basis for an intervention (Section 1.3).

1.2 Influence of Functional Impairment on Test Results

A functional impairment can influence the development of a child in areas other than the impairment itself. A visual impairment, for example, can influence different areas of development, such as concept development (Pagliano, 1998) and motor development (Getman, 1993), by hampering the interpretation of space and distance. In addition, visual impairment is often caused by cerebral factors which in many cases cause impairment in other areas of functioning as well. As a consequence, a large percentage of children with visual impairment have additional disabilities, such as intellectual disability (Teplin, 1995). This means that a visual impairment can lead to a lower score on an assessment of cognitive or motor development than the child would have achieved if he/she had not suffered from the visual impairment. In this case, the score adequately reflects the developmental level. However, the visual impairment can also influence the test results via another route. If test items require visual ability, although visual ability is not the target of the measurement, a child with visual impairment might not be able to show his or her abilities. This latter form of influence on the part of the impairment leads to test results that do not adequately reflect the developmental level of the child and are thus invalid.

The same can happen in cases of other types of impairment, such as motor impairment or speech/language impairment. A motor impairment can hamper cognitive development because it limits the child's exploration of the environment. However, if an instrument assesses the cognitive development of a child using jigsaws, for example, a motor impairment can limit the child's ability to show his or her cognitive abilities, which will lead to an underestimation of the actual cognitive abilities the child has.

One possible solution to the latter threat to validity is accommodating the assessment instruments in order to minimize the unwanted influence of impairments on test results. "Accommodating" means that changes are made to the assessment instrument with the aim of minimizing the influence of the impairment on the test results, without changing what the instrument measures (Alant & Casey, 2005; Batshaw Claire, Church, & Batshaw, 2007; Bondurant-Utz, 2002; Thurlow, Elliot, & Ysseldyke, 2003). The aim of accommodations is thus to lower the influence of the impairment on the test results by, for example, removing the visual or motor components from the test items for as much as possible. Accommodations can be made to the presentation format, response format, setting of a test, and timing of a test (Thurlow et al., 2003).

Some developmental assessment instruments, like the Stanford-Binet Intelligence Scales, Fifth Edition (Braden & Elliot, 2003; Roid, 2003), provide guidelines for accommodations in cases of special needs. The manual of the Bayley-III describes possible accommodations in cases of visual, hearing, or motor impairment (Bayley, 2006a). However, the accommodations described in the manual are not standardized: They have not been described extensively per item, and, consequently, it is up to the test administrator to decide how exactly to implement the accommodations. In addition, the accommodations described are in many cases not sufficient to reach the ultimate goal of a fair developmental assessment.

1.3 Assessment as a Basis for Intervention

The second problem associated with standardized developmental assessment instruments, which is considered in this thesis, relates to the assessment's purpose. Different purposes for developmental assessment have been identified, among which are diagnosis and indication analysis, intervention planning, and program evaluation (Neisworth & Bagnato, 2004; Snow & Van Hemel, 2008). The results of a developmental assessment can ideally be used for multiple purposes at the same time. Standardized developmental assessment instruments are necessary for diagnosis and indication analysis, but they generally do not estimate learning capacities and sensitivity to instruction. Consequently, they do not yield information that can be used as a basis for intervention planning (Bagnato, Neisworth, & Pretti-Frontczak, 2010; R. J. Kahn, 2000; Snow & Van Hemel, 2008). The information that can be used for a subsequent intervention is, however, often regarded as the most important part of the assessment results. Starting from the model of "Intervention-oriented assessment" ("Handelingsgerichte diagnostiek;" Pameijer & Van Beukering, 2004), which is widely applied in the Netherlands, the diagnostic process in early intervention practice should primarily be focused on yielding information for the intervention.

One possible solution is to combine the assessment of a standardized instrument with a dynamic assessment approach. Dynamic assessment includes a training phase to improve the child's performance on the test, and to identify the amount and type of assistance that a child needs to accomplish the test tasks (Haywood & Lidz, 2007; Resing, 2006). This provides information about the learning needs of a child and is therefore particularly valuable for intervention planning.

1.4 Aim and Outline of the Thesis

The research carried out as part of the current thesis aims to solve the problems described in Sections 1.2 and 1.3, and thereby enhance the suitability of the Bayley-III-NL for different subgroups of young children in early intervention. The following research questions form the basis for the research carried out:

- Which instruments for developmental assessment of young children are available and what is known about their suitability in cases of special needs?
- How can we accommodate the Bayley-III-NL so as to increase the suitability for children with visual impairment?
- How can we accommodate the Bayley-III-NL so as to increase the suitability for children with motor impairment?
- How can we accommodate the Bayley-III-NL so as to increase the suitability for children with speech/language impairment?
- How can we adjust the Bayley-III-NL so that the test results become more useful as a basis for intervention planning?

The research was funded by grant 157013002 of the Netherlands Organization for Health Research and Development (ZonMw). We collected the data in cooperation with 47 different branches of organizations supporting young children with special needs in the Netherlands. Five research papers form the basis for the current thesis. Figure 1.1 summarizes the content of the chapters and the structure of this thesis.

Figure 1.1 Outline of the Chapters

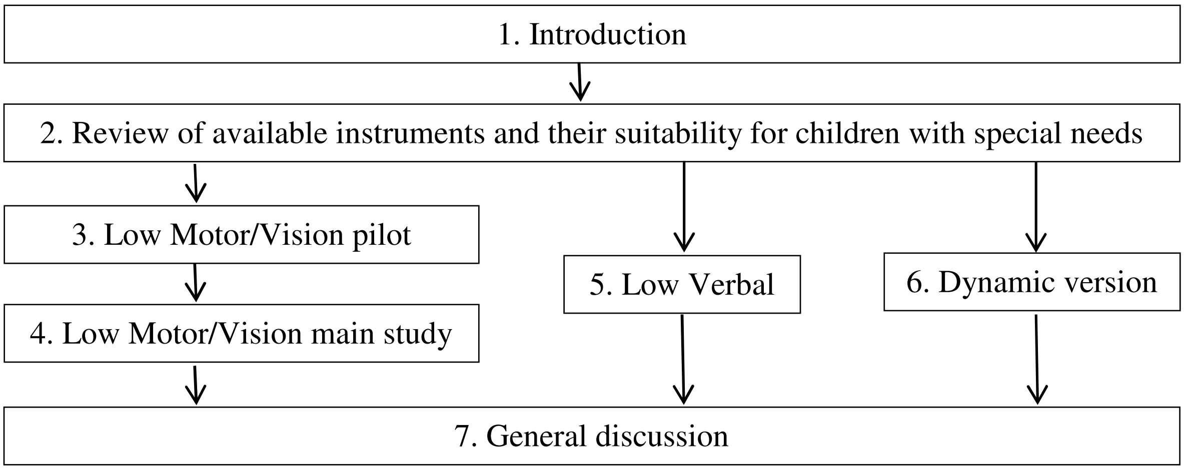

Figure 1.1 The arrows show the relationship between the chapters.

Chapter 2 describes the results of a literature review focusing on the suitability of standardized developmental assessment instruments for young children with special needs. It starts with a description of frequently encountered problems, followed by a systematic literature review, resulting in a list of currently available developmental assessment instruments for young children. We discuss the suitability of these instruments for children with different kinds of special needs.

The focus of **Chapter 3** is a pilot study on the “Low Motor/Vision” version of the Bayley-III-NL. To develop this version, we applied the concept of accommodation to the materials, item instructions, and procedures of the Bayley-III-NL. The aim was to decrease the influence of visual and motor impairment on the Bayley-III-NL test results and thereby increase the suitability of the instrument for testing young children with motor and/orvisual impairment. The accommodations were based on those of the “Low Motor” (Ruiter, Nakken, Van der Meulen, & Lunenborg, 2010) and “Low Vision” (Ruiter, Nakken, Janssen, Van der Meulen, & Looijestijn, 2011) versions of the Dutch Bayley Scales of Infant Development, Second Edition (BSID-II-NL; Bayley, 1993; Van der Meulen, Ruiter, Spelberg, & Smrkovsky, 2002), complemented with accommodations to the items that were new to the Bayley-III-NL. We were able to combine the Low Motor and Low Vision accommodations into one version for the Bayley-III-NL. This has the advantage that the accommodations can easily be combined when children with both motor and visual impairments are assessed. The chapter describes in detail the accommodations and the target group, as well as the results of the pilot study.

Chapter 4 describes the results of the main study on the Low Motor/Vision version of the Bayley-III-NL. In this study, we assessed a larger group of children with motor and/or visual impairment using the Bayley-III-NL Low Motor/Vision version as well as the standard version. The study also included a control group of children without impairment. This allowed us to not only answer the question whether children with impairment appear to benefit from the accommodations but also the question whether the content and difficulty of the test items remained the same in spite of the accommodations.

The focus of **Chapter 5** is the “Low Verbal” version of the Bayley-III-NL. The target group of this version is young children with speech/language impairment. The chapter briefly describes the results of a pilot study, and is then followed by a description of the main research. In the main research we compared the test results of children from the target group using the Low Verbal version to those of a large group of children without any impairment using the standard version of the Bayley-III-NL. In addition, we evaluated the added value of the Low Verbal accommodations, using an evaluation form and expert interviews.

Chapter 6 describes the results of a study on the dynamic version of the Bayley-III-NL. After providing an overview of the theoretical background of dynamic testing, we describe in detail the dynamic version of the Bayley-III-NL, followed by a description of the test results of a group of children with developmental disabilities using the dynamic version, along with the results of interviews with experts.

Chapter 7 is a general discussion of the results of the studies described in the light of the research questions, followed by the most important limitations of the studies, and the implications of the studies’ results for early intervention practice. This thesis concludes with suggestions for future research.

All references have been merged into one reference list, which can be found immediately following Chapter 7.

Chapter 2

A review of standardized developmental assessment instruments for young children and their applicability for children with special needs*

* This chapter has been published as: Visser, L., Ruiter, S.A.J., Van der Meulen, B.F., Ruijssenaars, A.J.J.M., & Timmerman, M.E. (2012). A review of standardized developmental assessment instruments for young children and their applicability for children with special needs. *Journal of Cognitive Education and Psychology, 11*, 102-127. doi: 10.1891/1945-8959.11.2.102

Abstract

This article provides a review of contemporary instruments for the developmental assessment of children aged 0 – 4 years and their applicability for children with special needs. The issues involved in the developmental assessment of children with special needs are discussed, and, on the basis of these issues, various instruments are then evaluated. *Method:* A literature search was carried out for articles about or using standardized developmental assessment instruments for children aged 0 – 4 years. *Results:* Eighteen instruments were found, of which 2 were nonverbal and 2 were designed for motor-impaired children. The instruments varied in terms of their suitability for children with special needs. *Conclusion:* The range of instruments is limited, especially for children younger than 2 years. Instruments for children with motor or hearing/language impairments are available, but their psychometric properties need to be researched and improved. For children with a visual impairment, no appropriate instrument is currently available.

2.1 Introduction

Good assessment practices are widely recognized as a key component of high-quality early childhood intervention (American Academy of Pediatrics, 2001; National Association for the Education of Young Children & National Association of Early Childhood Specialists in State Departments of Education, 2003). Appropriate methods for the developmental assessment of young children are those that correctly document the broad range of skills that emerge and develop during the first years of life. Standardized instruments are essential because they are the only tools capable of obtaining objective quantitative information about development directly from the child. Although doubt has been cast on its predictive value (Petermann & Macha, 2008), early assessment still appears to be valuable for those young children who achieve very low scores on assessment instruments (Gregory, 2007). The correlation with later IQ among the full population of children is in fact reasonable and much stronger than among the group of children who achieve an average score (Gregory, 2007).

Early assessment is essential for children with special needs because special needs limit children's ability to explore and understand the world around them and increase the chance of developmental delay (Best & Corn, 1993). Early detection of developmental problems is needed to facilitate an early start in support for parents and children, which has proven to be beneficial (Guralnick & Conlon, 2007; Shonkoff & Phillips, 2000) and cost-effective (Rémillard, 2008) by minimizing the long-term effects of developmental problems. When evaluating the development of a child with special needs, cognitive, language, and motor abilities should be evaluated independently from physical limitations or limitations because of the procedure and structure of the instrument. Because of the great impact test results can have on a child's life and that of the child's family, the utility and technical adequacy of developmental assessment instruments are often the subject of research.

The subject of this article is the use of standardized developmental assessment instruments with children aged 0 – 4 years with special needs. We will be distinguishing three major categories of adjustments for making an instrument suitable for use of children with physical impairments: adjustments for motor impairments, adjustments for visual impairments, and nonverbal adjustments (to increase the suitability for children with hearing or speech/language impairments). In accordance with Snow and Van Hemel (2008), we will define developmental assessment as the ongoing process of observing a child's current skills and using that information to help the child develop further within the context of the family and caregiving and learning environments. Usually, developmental assessment comprises the use of a combination of instruments, among which are the parental report, observation, and qualitative instruments. When information is desirable about the level of development in relation to peers, an important component of the early assessment process is the administration of an individualized, standardized, and norm-referenced developmental assessment instrument. To increase readability, we will refer to such instruments in general terms as simply "standardized instruments."

An important assumption underlying standardized developmental assessment instruments is that of an ordinal development of skills (Petermann & Macha, 2008). This implies that standardized instruments can only be sensibly used for children with special needs as long as they undergo a development that is only quantitatively, and not qualitatively, different from typically developing children. This assumption is also called the "similar sequence hypothesis" and means that "retarded and nonretarded persons traverse the same stages of cognitive development in the same order, differing only in the rate at which they progress and in the ultimate developing ceiling they attain" (Weisz & Zigler, 1979, p. 831). The tenability of this hypothesis has been partly supported by Kahn (1976) and O'Connell (1994) for children with profound intellectual and multiple disabilities. However, other authors presume qualitative differences in development (Orelove & Sobsey, 1996; Van der Putten, Vlaskamp, Reynders, & Nakken, 2005). Moreover, qualitative differences have been found in the fine and gross motor development of blind children (Brambring, 2007; Vink, 1994).

An alternative or addition to standardized instruments is the use of qualitative assessment instruments, which do not assume successive developmental stages. These instruments lack norms and usually lack evidence of reliability and validity (Hebbeler et al., 2008). However, the qualitative information yielded by such instruments can be highly valuable in clinical practice, depending on the aim of the assessment. Qualitative instruments, however, are not included in this review.

Some authors (Bagnato, Neisworth, & Pretti-Frontczak, 2010; Neisworth & Bagnato, 1992; 2004; Snow & Van Hemel, 2008) argue that standardized instruments should not be used in the early assessment process because of certain problems that we will discuss in detail further on. Others, such as Flanagan and

Alfonso (1995) and Allen (2007), reason that standardized instruments are essential, because they provide information that is not provided by other kinds of instruments. Standardized instruments are the only tool for obtaining objective quantitative information about the child's development directly from the child. Another argument for using standardized instruments is that many countries (e.g., the Netherlands [Pijl & Hamstra, 2005] and Europe in general [Parveva, De Coster, & Noorani, 2009]) incorporate a quantitative component in their special education eligibility criteria for preschool children.

Although essential, standardized instruments do have limitations, especially when applied to the assessment of children with special needs. What follows is an overview of what has been written in the literature about the problems encountered and possible solutions for them, divided into four topics: reliability across levels of developmental functioning, dependence of test results on specific skills, information that is supportive in developing intervention strategies, and test duration. These issues should be taken into account when making a choice for a particular instrument because instrument characteristics vary in relation to the limitations.

2.1.1 Issues In The Standardized Assessment of Children with Special Needs

1. Reliability across levels of developmental functioning. As a consequence of the characteristics of the research sample as well as the methods used for developing norms, reliability in the low range of norm score distributions is often problematic (Meisels & Atkins-Burnett, 2000). This is especially true for children younger than the age of 4 years and even more so if they have a developmental delay (Bracken, 2000). Low reliability is a consequence of high variability in scores between testees of comparative performance level (Albers & Grieve, 2007) and is manifested in broad reliability intervals.

Low reliability can become evident through problems with test floors and item gradients. When the test floor is problematic, it is not possible to obtain a score below a certain limit, or low test scores are not sufficiently reliable. This is caused by characteristics of the standardization sample, which is often composed of relatively many children obtaining an average score. Flanagan and Alfonso (1995) conclude that, with a few exceptions, test floors of preschool intelligence tests are poor. An item gradient refers to "the steepness with which standard scores change from one raw score unit to another" (Rathvon, 2004, p. 49). Bracken (1987) suggests the minimal requirement of no fewer than three raw score items per standard score standard deviation. Flanagan and Alfonso (1995) use this criterion and report that the majority of preschool intelligence tests do not have adequate item gradients across all levels.

A relatively easy way to minimize the test floor and item gradient problems without lengthening the test is to increase the number of scoring alternatives. A polytomous scoring system has been shown to result in improvements in test scores for children who do not pass an item but have the potential to acquire the skill soon. It also enables a better differentiation to be made within the group of children who do not pass an item, resulting in improved reliability (Tzuriel, 2000).

2. Dependence of test results on specific skills. The suitability of instruments for children with special needs is hindered by the fact that the procedures of most standardized instruments are developed for a smooth administration with typically developing children. Standardization research is done with typically developing children as well (Neisworth & Bagnato, 2004). For the purpose of maintaining standardization and application of the norm tables, examiners cannot deviate from the administration procedures dictated in the test manual. Consequently, a child with special needs might be unable to show his or her abilities in the field of interest. A child with a fine motor impairment, for example, might not be able to build a tower of blocks because the blocks are too small, whereas the task is actually meant to be a measurement of cognitive abilities, not motor ones. This yields invalid test results (Ruiter, 2007). When the child's disability has an unwanted influence on the test result, the test procedures become inadequate and less valid (Skovgaard, Houmann, Landorph, & Christiansen, 2004). It is striking that the existing instruments seem to be unsuitable for children with special needs, because it is this group of children that is particularly in need of developmental assessment.

Adaptations to standard procedures are thus necessary. The manuals of some standardized instruments describe possible accommodations; for example, placing objects in the child's hands, or using demonstrative gestures (Bayley, 2006a). However, these accommodations are usually not described in detail. If detailed guidance is provided, research on the influence of the accommodations on the test results is often lacking (Hebbeler et al., 2008). This is problematic because the accommodations might influence the items in such a way that the difficulty or even the concept they measure is changed.

The solution is the development of instruments specifically designed for and studied among children with special needs, or the standardized accommodation of materials and procedures of instruments that, in their original form, are not specifically designed for a special group of children. In this regard, the information that is given in some manuals on test results with special groups of children is insufficient. The children in almost all special groups obtain a below average score, but no information is provided on whether the children have been able to show what abilities they have. In other words, the results on special group studies do not sufficiently support the idea that the instrument can be used for an accurate and valid developmental assessment of children in the special groups.

3. Information supportive for developing intervention strategies. Early assessment has various purposes. Generally, the following categories are distinguished: screening, diagnosis and eligibility determination, individualized program planning, and program evaluation (Neisworth & Bagnato, 2004; Snow & Van Hemel, 2008). Screening can be seen as a separate purpose because it takes place at a different stage in the assessment and with a wider target group than assessment for other purposes. Specific screening instruments are available (Glascoe, 2005). Some authors argue that for individualized program planning, standardized instruments should not be used (Bagnato et al., 2010; Snow & Van Hemel, 2008).

However, it would be highly desirable for clinical practice to have a single assessment instrument that can be used for all purposes simultaneously, because diagnosis, eligibility determination, and program planning often overlap in time. This would save time and money and reduce the impact on the child. However, norm-referenced assessment items are selected based on the ability to discriminate among children. As a consequence, the items often measure nonfunctional skills that appear not to be useful as a basis for intervention (Neisworth & Bagnato, 2004).

The extent to which assessment results can be used for setting up interventions depends on the assessment process, the test(s) used, and the abilities of the assessor in linking assessment to intervention. Regarding the assessment process, information from different sources, measurements, and contexts must be obtained, and the process must be an evolving and cyclic one (Conlon, 2002; Emde & Robinson, 2000; Meisels & Atkins-Burnett, 2000; Romanczyk et al., 2005). Priorities for the child and the child’s family should be assessed in addition to the child’s skills so that intervention can be attuned to the child’s needs (Whinnery & Whinnery, 2007). Regarding the test requirements, the test results should provide information about a child’s competencies and resources as well as delays because an intervention that is attuned to the child's strengths gives the best results (Conlon, 2002; Provence, Erikson, Vater, & Palmeri, 1995). Interpretation is enhanced if a test manual provides extensive information about the content of the subscales (Kaufman & Kaufman, 2004) and, ideally, about the inferred consequences of particular test results for intervention. Test results also need to provide information on how a child manifests a particular skill (Meisels, 1994) and on the learning potential of the child (Tzuriel, 2000). The aim of an intervention is usually to change behavior, in other words, learning. Information on the learning potential in different areas will therefore help in deciding in which areas an intervention will be most effective. Insight into the learning potential of a child is also expected to enhance the prediction of future outcome (Tzuriel, 2000). The use of polytomous scoring enhances the usefulness of test results because budding skills can be identified by being scored as "partially acquired." These skills can be selected as being good candidates for intervention (Hoekstra, Jansen, Van der Meulen, Oenema-Mostert, & Ruijssenaars, 2010). This is related to the theory of Vygotsky (1978) concerning the Zone of Proximal Development in which a child can reach a developmental level that is higher than the Zone of Actual Development with the help of a more capable person. The Zone of Proximal Development should thus be the focus of intervention.

The test requirements mentioned are relevant in the context of this article. It is important to keep in mind that the actual usefulness of test results for the intervention depends on the assessment process and individual assessor qualities.

4. Test duration. Standardized assessment procedures, as used for older children and adults, are inappropriate for young children, because young children cannot focus for the same length of time (Hebbeler et al., 2008; Rolfe, 1994). Yet, such standardized assessment instruments typically have long test durations. This becomes a problem when testing young children and even more so when the children have special needs, because children with special needs generally need more time for processing an instruction and completing a task.

A way of reducing test duration is standardized adaptive testing, a method to achieve short test durations while guaranteeing a minimum of reliability. Items are administered depending on the scores on previous items. This can be done using item response theory and on a computerized basis (Van der Linden, 2008). It has the potential of shortening an instrument and optimally adapting the instrument to the child while obtaining an optimal amount of information about development and not harming standardization. The use of floor and ceiling rules is also a means of adaptive testing and is incorporated in many of the existing instruments.

In this article we will present an overview of widely used standardized developmental assessment instruments and their characteristics as related to their use with children with special needs. The four issues described will be used as a basis for the review. We expect this overview to be useful for researchers as well as clinicians. It may help in making a choice for a particular standardized instrument based on information about the strengths and weaknesses of current instruments, instead of choosing an instrument only based on availability and maybe just habit.

Next, we will describe the methods used in this review, including the target group and the literature search. The results form an overview of widely used instruments for the developmental assessment of young children and their characteristics in relation to their use with children with special needs. This article represents an update of earlier reviews of infant developmental assessment instruments (Bracken, 1987; Bradley-Johnson, 2001; Flanagan & Alfonso, 1995). However, although these reviews focused on the psychometric characteristics of the instruments, the focus in this article will be on the applicability of the instruments for children with special needs.

2.2 Method

2.2.1 Target Group

The target group of the instruments in this review was children younger than 4 years of age. The target group of our review is young children with special needs, including developmental delay, disorders such as autism or attention deficit/hyperactivity disorder (ADHD), language problems, and visual, hearing, or motor impairments.

2.2.2 Literature Search

We used a literature search of the established databases PsychInfo and Eric to find high-quality and widely used instruments that are currently being used in the developmental assessment of young children. We used the following key words:

- “assessment” or “test” or “measurement” or “batter*”
- “young children” or “infant(s)” or “preschool(er)(s)”
- “standardised” or “standardized” or “standardisation” or “standardization.”

These combinations of words were merged in a combined search using the Boolean operator “AND.” Using “*batter**” we searched for the terms “battery” and “batteries.” All words were searched for in the title and/or abstract. Only peer-reviewed articles published in 1995 or later were selected. The literature search was done using English keywords only. It resulted in a list of 399 publications.

From this list, we selected those articles discussing or using one or more standardized instruments with a target group that included children younger than the age of 3 years. Instruments with an age range starting at 3 years were not included. Including these would have added a great many instruments for a broad age range, including adults, which did not have young children as their main target group. Instruments were included in the review if they were published 15 years ago or less, as indicated by the publication date, because test norms older than 15 years are not accurate as a result of the Flynn effect (Flynn, 1984).

We excluded those instruments measuring skills in only a single area and neonatal scales. For a review of instruments assessing motor development in infants, we refer the reader to Heineman and Hadders-Algra (2008) and for a review of preschool language scales to Friberg (2010). We also excluded screening instruments (as indicated by the name of the instrument or its described main purpose) and parental reports. For an overview of screening instruments and their characteristics, see Glascoe (2005).

After this selection, 89 articles remained, which we included in the review. Altogether, these articles reviewed, mentioned, or used 18 different standardized instruments. Only the latest version of the instruments was included. A reference list of the 89 articles is available from the first author upon request.

To assess the qualities of the 18 instruments, we looked up test reviews by searching for the instrument’s name in handbooks and in the databases PsychInfo, Eric, and Google Scholar. All instruments were reviewed in light of the four issues described in the introduction. The evaluation was based on the information and judgments found in the reviews of the instruments. The criteria were as follows:

1. Reliability across levels of developmental functioning. Reliability was estimated based on the internal consistency measurements. When reliability values were given, they were judged as high (≥ .90), moderate (.80 – .90), or low (<.80), in accordance with Flanagan and Alfonso (1995). The evaluation of the test floors and item gradients of the instruments was based on the information and judgments found in the reviews of the instrument. We gave the index score to specify the test floor. Most

instruments use an index score, which is normed such that, in the target population, the mean is 100 and the standard deviation is 15. When no information at all was available in published articles or handbooks, this was indicated by the abbreviation “ni.”

2. Dependence of test results on specific skills. We described what research has been done concerning special groups. This does not serve as a support for the suitability of the instrument for the groups examined. It does, however, provide information about which special groups the instrument is intended to be suitable for and what the average test results were. We also described what accommodations were made available for specific groups of children.
3. Information that is supportive in developing intervention strategies. We indicated whether the instruments explicitly intend to provide information that is relevant for intervention planning, in addition to a global score, and in what way.
4. The test duration is given in minutes, when available. As far as we know, no research has been done on test duration with children younger than 4 years of age. Based on our own experience, we consider the test duration to be adequate if it is an hour or less. If the duration exceeds an hour, the child is generally too tired and unfocused for a reliable measurement to be obtained.

2.3 Results

In Table 2.1, an overview is presented of the main characteristics of the 18 instruments, including the age range, country(-ies) and size(s) of standardization sample(s) reported, and the numbers of articles found in the literature search that reported research that have applied the instruments. The order of the instruments is based on the number of articles found in our search (from high to low).

In the following section, we will discuss the characteristics of the instruments in view of the assessment of children with special needs. The numbers of each of the instruments refer to the numbers in Table 2.1. In addition to information related to the issues described in the introduction, for each instrument we will describe which area(s) of development it assesses and in which languages the test is available. The characteristics described will then be summarized in Table 2.2.

We would like to emphasize that this is not an extensive review of the psychometric properties of these instruments but an overview of what is known about each instrument's characteristics in relation to its applicability in the assessment of children with special needs.

Table 2.1
Overview of Standardized Developmental Assessment Instruments for Young Children

No. / Name	Author(s), Year	Age Range (Years; Months)	Standardization Sample (n, Country)	No. of Articles
1. Bayley-III	Bayley, 2006	0;1 – 3;6	1.700, US	46
2. Stanford-Binet 5th Ed.	Roid, 2003	2;0 – 85+	4.800, US	12
3. WPPSI-III	Wechsler, 2002	2;6 – 7;11	1.700, US	11
	Wechsler, 2003		805, UK	
	Wechsler, 2003		700, Canada	
	Wechsler, 2004		999, France	
	Wechsler, 2004		545, Australia	
	Wechsler, 2005		UK sample, Sweden	
	Ricken et al., 2007		710, Germany	
	Wechsler, 2008		246, Norway	
	Wechsler, 2008		987, Italy	
	Wechsler, 2009		714, Finland	
	Wechsler, 2009		1.220, Spain	
	Wechsler, 2009		1.672, Neth	
4. Kaufman-ABC-II	Kaufman & Kaufman, 2004	3;0 – 18;11	3.025, US	8
	Melchers & Preuß, 2009		3.098, Germany	
5. Woodcock-Johnson-III	Woodcock et al., 2001	2;0 – 90+	8818, US	7
	Wechsler et al., 2010		1094, Brazil	
6. BDI-2	Newborg, 2005	0;0 – 8;0	2500, US	6
7. Griffiths 0-2	Griffiths & Huntley, 1996	0;0 – 2;0	665, UK	5
8. BAS-II /	Elliot, 1996/	2;6 – 17;11	1.689, UK	5
9. DAS-II	Elliot, 2007		3.480, US	
10. PEP-3	Schopler et al., 2005	0;6 – 7;0	555, US	4

(Table continues)

Table 2.1 ***(continued)***

No. / Name	Author(s), Year	Age Range (Years; Months)	Standardization Sample (n, Country)	No. of Articles
11. SON-R 2½-7	Tellegen et al., 1998	2;6 - 7;0	1.124, Neth	4
	Tellegen et al., 1998		Neth sample, UK	
	Tellegen et al., 2007		1.027, Germany	
	Tellegen et al., 2008		450, Czech	
	Tellegen et al., 2009		Neth sample, France	
	Tellegen et al., 2009		252, Slovakia	
	Laros et al., 2011		120, Brazil	
12. Griffiths 2-8	Luiz et al., 2006	2;0 – 8;0	1.026, UK	3
13. ET 6-6	Petermann et al., 2006	0;6 – 6;0	950, Germany	1
14. DAYC	Voress & Maddox, 1998	0;0 – 5;11	1.269, US	1
15. Brigance IED-II	Brigance, 2004	0;0 – 7;0	1.156, US	1
16. Leiter-R	Roid & Miller, 1997	2;0 – 20;0	1.719, US	1
17. BSID-II-NL-Low Motor	Ruiter et al., 2009	1;0 – 3;6	1.700, Neth	1
18. MMFC	Mayes, 1999	0;0 – 2;0	na	1

Note. na = not available; Neth = Netherlands; US = United States; UK = United Kingdom.

1. The *Bayley Scales of Infant and Toddler Development, Third Edition* (Bayley-III; Bayley, 2006a) is a widely used instrument. As many as 46 articles were found that discussed or used the Bayley-III, as shown in Table 2.1. It is an English language test developed in the United States and assesses the areas of cognition, receptive and expressive communication, along with fine and gross motor development, social-emotional development, and adaptive behavior. The psychometric properties are generally good, but low reliability coefficients (.71) were obtained in the younger age groups (1-5 months) within the Receptive and Expressive Communication subtests (Albers & Grieve, 2007). For the other subtests and age groups, the coefficients range between .72 and .98, with an average of 0.89 (Bayley, 2006b). The test floor is an index score of 55. Item gradients are problematic, especially in the lowest age groups (Bayley, 2006a). Test results are provided for children in the following clinical samples: Down syndrome, pervasive

developmental disorder, cerebral palsy, language impairment, developmental delay, asphyxiation at birth, small for gestational age, and premature birth or low birth weight. The children from the special groups score lower than children from the control group (Bayley, 2006b). The manual gives an overview of possible accommodations to the standard test procedure, which are intended to minimize the effects of motor involvement. The instrument cannot be used to obtain a norm-referenced score for a severely impaired child (Bayley, 2006b). No evidence is presented on predictive validity and accuracy or how the intervention provision is improved as a result of a Bayley-III administration (Albers & Grieve, 2007). The administration time is 90 minutes for children aged 13 months and older.

2. The *Stanford-Binet Intelligence Scales, Fifth Edition* (Roid, 2003) consists of the scales of fluid reasoning, knowledge, quantitative reasoning, visual-spatial processing, and working memory. Ward, Rothlisberg, McIntosch, and Bradley (2011) found that results on the tests for preschool children can best be interpreted in terms of an overall ability model and not by distinguishing separate factors. Although the first version of the test was developed in France, the newest version is only available in English and was developed in the United States. Reliability is moderate to high across ages, with subtest internal consistency values ranging from .84 to .89 and for the factors from .90 to .98 (Bain & Allin, 2005). There is no problem with the test floor because it is possible to extend low-end items (Becker, 2003). Information on item gradients is not available. Information about applications with special populations (mental retardation, developmental delay, autism, speech/language disorders, learning disabilities, and motor impairments) is fairly comprehensive. Accommodations are described extensively (Braden & Elliot, 2003), but a description of the accommodations used during the special groups research for the group with motor impairments is not provided (Bain & Allin, 2005). The instrument has a nonverbal component, requiring the examinee to point, make movement responses, or assemble manipulatives, but a minimum of receptive and expressive language skills is still required. Although the instrument intends to provide information for planning intervention, studies supporting intervention claims are not included in the technical manual (Bain & Allin, 2005). Becker (2003) judges the administration time for younger children as acceptable. The administration time is reduced by an adaptive testing procedure, in which examiners use the information they have about an examinee to determine where to begin testing and to select only those items that are appropriate for that examinee.

3. The *Wechsler Preschool and Primary Scale of Intelligence, Third Edition* (WPPSI-III) is composed of the scales of receptive vocabulary, information, block design, picture completion, and picture naming for the youngest age group (2 years and 6 months to 3 years and 11 months). This newest version is available in 12 countries: United States (D. Wechsler, 2002a), United Kingdom (D. Wechsler, 2003a), Canada (D. Wechsler, 2003b), France (D. Wechsler, 2004a), Australia (D. Wechsler, 2004b), Sweden (D. Wechsler, 2005), Germany (Ricken, Fritz, Schuck, & Preuß, 2007), Norway (D. Wechsler, 2008a), Italy (D. Wechsler, 2008b), Finland (D. Wechsler, 2009a), Spain (D. Wechsler, 2009b), and the Netherlands (D. Wechsler, 2009c). A total IQ, performance IQ, verbal IQ, and a general language composite can be calculated based on the test scores. Internal consistency reliability coefficients for the youngest age group range from .45 to .83 for the subtests and from .73 to .87 for the composites (D. Wechsler, 2009c). The test floor is 55, but there is a limited breadth of coverage for children in the lowest age group, along with a limited range of scores for children who are extremely low or high functioning (Sattler, 2008). The norm tables show some problems with item gradients (D. Wechsler, 2009c). The validity for different clinical groups (mental retardation, developmental delay, ADHD, cognitively gifted, autism, and language disorder) has been studied. The average score for children in these groups is below the population average, except for the gifted children, but no information is provided on whether the test results are valid (D. Wechsler, 2002b). According to Sattler (2008), the instrument is useful for children with mild physical disabilities. One of the characteristics that make the instrument suitable is that the test administrator is allowed to repeat the questions as often as necessary and to encourage the child. The manual not only gives brief guidelines (accommodations) for testing children with special needs, but also mentions that the test might yield scores that underestimate the ability of children with sensory or motor impairments. The manual does not mention how test results can be used for intervention planning. The administration time for the youngest age group is 35-42 minutes (D. Wechsler, 2009a; 2009b; 2009c).

4. The *Kaufman Assessment Battery for Children, Second Edition* (Kaufman-ABC-II; Kaufman & Kaufman, 2004) is an English-language instrument designed for children aged 3 years and older. The previous version had a starting point of 2 years and 6 months and was adapted as the Dutch “Groninger Ontwikkelingsschalen” (GOS 2½-4½; Neutel, Van der Meulen, & Lutje Spelberg, 1996). The second edition is also available in German (Melchers & Preuβ, 2009).

Subtests for the youngest age group include word order, conceptual thinking, face recognition, triangles, expressive vocabulary, and riddles, but only the global score was found to be valid (Bain & Gray, 2008). In the study of the Dutch version, the original factor structure could not be replicated (Neutel et al., 1996). Internal consistency reliability coefficients for ages 3- 6 years range from .69 to .91 for the subtests and from .90 to .95 for the scales “sequential,” “simultaneous,” “learning,” “planning,” and “knowledge,” and for the Mental Processing Index, Fluid-Crystallized Index, and Nonverbal Index. Only 3 out of 17 subtests have a reliability coefficient below .80. In contrast to the previous version of the instrument, this second edition has an adequate test floor for preschool children (Kaufman, Lichtenberger, Fletcher-Janzen, & Kaufman, 2005). Information on item gradients is not available. Clinical validity studies have been carried out for children with specific learning disabilities, mild mental retardation, autism, ADHD, hearing loss, and gifted children. These studies yielded below average scores except for the gifted children who obtain above average scores. However, the clinical validity studies do not provide any information about the tool’s validity for these groups of children. Built-in accommodations include a nonverbal scale and teaching items for nearly all subtests, designed to help children who initially respond incorrectly by explaining and demonstrating the correct response and allowing a second trial. This is thought to deliver additional information about the learning potential of the child, which is useful in setting up an intervention plan. For the younger children, the administration time is around 25 minutes (Bain & Gray, 2008).

5. The Woodcock-Johnson III (Woodcock, McGrew, & Mathner, 2001a; 2001b) is a cognition and achievement instrument. The original version is in English, and there is a Spanish version as well (S. M. Wechsler et al., 2010). The cognitive subtest consists of the categories of verbal ability, thinking ability, and cognitive efficiency. In addition, there are several clinical and achievement subtests. Reliability is adequate on a factor level, but internal consistency on the subtests ranges from .76 to .97 (Sattler, 2001). The test floors are appropriate only for children aged 2 years and 8 months or older, depending on the subtest (Tusing, Maricle, & Ford, 2003). Bradley-Johnson and Durmusoglu (2005) reviewed the reading and math subtests of the Woodcock-Johnson III and found that the item gradients were inappropriate because in some cases, one point change in raw score changed the index scores by as many as 15 points. The manual provides information about the non-standardized accommodations that are made possible so as to allow individuals with special needs (including young children, individuals with learning

or reading difficulties, with attention or behavioral difficulties, and with hearing, visual, or physical impairments) to participate more fully in the process (Blackwell, 2001). The link with intervention is facilitated by the aim of the instrument, which is to reliably differentiate children's abilities in terms of specific predictive domains that are related to early learning (Tusing et al., 2003). The administration time is 45-50 minutes for the standard battery and 1½-1¾ hours for the extended battery (Blackwell, 2001).

6. The *Battelle Developmental Inventory, Second Edition* (BDI-2; Newborg, 2005) is criterion-referenced as well as norm-referenced and can be administered in English and Spanish. The instrument is easy to administer. The reliability is moderate to high for the total score and the domains but for several subdomains, the internal consistency coefficients fall well below the recommended .80 (Bliss, 2007). The test floor is an index score of 40. Bliss (2007) concludes that the item gradients are acceptable, but writes that a raw score increase of 1 point results in no more than a 2-point increase in scaled scores, which implies that the item gradients do not meet the criteria as set by Bracken (1987). The manual indicates that the instrument is suitable for children with autism, developmental delays, motor delays, speech and language delays, and for premature children. Accommodations for children with disabilities are provided, but it is unclear how these affect the scores obtained (Bliss, 2007). The manual mentions planning and providing instruction (i.e., intervention) as one of the specific purposes of the test, but no specific attention is given as to how the assessment results can be used for this purpose. Administration time is 60-90 minutes.

7 & 12. The *Griffiths Mental Development Scales – Revised: Birth to 2 Years* (Griffiths & Huntley, 1996) and the *Extended Revised: 2 to 8 Years* (Luiz et al., 2006) versions are standardized in the United Kingdom. The instrument consists of the following scales: locomotor, personal-social, hearing and language, eye and hand coordination, and performance. The extended version has an additional subscale measuring practical reasoning. Much research has been done on the reliability of previous versions of the instrument but not on the newest version (S. Johnson & Marlow, 2006; Luiz, Foxcroft, & Tukulu, 2004). Information on the item gradients is not available. The index score can be as low as 35 (Barnett et al., 2004), which indicates that the test floor is adequate. No validation research for special groups has been done and no accommodations are described for children with special needs. No special attention is paid to the use of test results in setting up an intervention. Administration time is 50-60 minutes.

8 & 9. The preschool edition of the *Differential Ability Scales, Second Edition* (DAS-II; Elliott, 2007) is the American version of the *British Ability Scales, Second Edition* (BAS-II; Elliott, 1996) and can be administered in Spanish and English. The two scales have very similar characteristics. The lower-level battery (early years, age 2 years and 6 months to 3 years and 5 months) of both instruments consists of four core subtests that generate verbal and nonverbal ability composite scores (Marshall, McGoey, & Moschos, 2011). The internal consistency reliability coefficients of the DAS-II are adequate for the early years, ranging as they do from .79 to .94. Because the reliability in Table 2.2 is judged based on the reliability across the full age range, the lowest coefficient of .79 results in a negative appraisal. Reliability coefficients derived from the special groups range from .42 to .99, with many coefficients lower than those reported for the normative sample. The test floor for the DAS-II is appropriate because of the extended norms provided (Beran, 2007). The BAS-II floor varies per subscale, with a maximum of 51 for the early years (Sparrow & Davis, 2000). The suitability for use with preschool children is facilitated by batteries that have been designed specifically for their particular age ranges rather than merely having been extended from school age to preschool age (Hill, 2005). However, children performing younger than the age level of 2 years and 6 months are unlikely to meet the basal requirements for most of the subtests, resulting in an inability to calculate meaningful standard scores (Klinger, O'Kelley, & Mussey, 2009). No information is available about the item gradients. According to Beran (2007), the test scores appear to be valid for children who are gifted, use sign language, have mental retardation, a reading and/or written expression disorder, a mathematics disorder, ADHD, ADHD plus a learning disorder, an expressive language disorder, a mixed receptive-expressive language disorder, limited English proficiency, or who are at risk. Examiners should be cautious when using the instrument with examinees with fine motor difficulties because several subtests require manipulation of materials, and there are no accommodations made available for this group. The nonverbal score makes the instrument appropriate for children with hearing, or speech and language impairments (Beran, 2007). There is no strict administration order for the items. This enables individualization of test sessions, which is an advantage, especially when working with children with special needs (Klinger et al., 2009). It is possible to collect only relevant and specific cognitive data, which supports a more purposeful assessment and usefulness for interventions, and reduces administration time. The administration time for the preschool version is 20-39 minutes (Beran, 2007; Klinger et al., 2009). This short

duration is caused by the use of age-related starting points, decision points, and alternative stopping points (Hill, 2005).

10. The *Psychoeducational Profile, Third Edition* (PEP-3; Schopler, Lansing, Reichler, & Marcus, 2005) was designed to assess the skills of children with autism spectrum disorder (ASD) and communicative disabilities. A Chinese version is under construction (Fu et al., 2010). The previous version (Psychoeducational Profile-Revised [PEP-R]) has been translated for use in Brazil (De Leon, Bosa, Hugo, & Hutz, 2004), Estonia (Kikas & Haïdkind, 2003), the Netherlands (Steerneman, Muris, Merckelbach, & Willems, 1997), and Italy (Villa et al., 2010), but no separate standardization research has been done. The English language standardization sample was made up of 407 children with ASD and 148 typically developing children. The instrument consists of 10 subtests belonging to the communication, motor, or maladaptive behavior scales. Internal consistency reliability coefficients for the subtests and composites range from .84 to .99. No information is provided on the test floor and item gradients. The test has a trichotomous scoring system. The PEP-3 was developed to be suitable for children with an autism spectrum disorder, communication problems, or developmental delay because the test items are presented using simple, concrete instructions and because most of the expected responses are nonverbal. There are no specific accommodations for children with other impairments. The instrument provides additional data about special learning strengths and teachable skills, which is expected to help in setting up intervention strategies (Naglieri & Chambers, 2009). Administration time is 45-90 minutes.

13. The *Developmental Test 6 Months to 6 Years* (ET 6-6; Petermann, Stein, & Macha, 2006) was developed for the assessment of cognitive, gross and fine motor, language, social, and emotional development in German-speaking countries. It is administered while playing with the child (Macha & Petermann, 2008). Internal consistency reliability has not been examined, and no information is given about the test floors or item gradients. No specific accommodations are described, but the instrument appears to be suitable for children with motor impairments (Macha, Mayer, Petermann, Petermann, & Waldeck, 2007). The validity when used with language-impaired children is limited because the instrument is language-oriented. No information is given about the information yielded for intervention. Administration time is less than 45 minutes on average (Macha & Petermann, 2008)(Macha & Petermann, 2008).

14. The *Developmental Assessment of Young Children* (DAYC; Voress & Maddox, 1998) is an English language test assessing cognition, communication, social emotional development, physical development, and adaptive behavior. Reliability coefficients range from .90 to .99 (Pro-Ed, Inc., 2011). The floor scores are inadequate and the manual states that one should rely on supplemental information when assessing children less than 12 months of age. No information is available about the item gradients. Test administration can be adjusted to the needs of the child by administering part of the subtests and by testing the child in the child's natural environment. However, the instrument is standardized with verbal instructions and no specific accommodations are described for use with children with special needs (Voress & Maddox, 1998). A positive factor for how test results relate to intervention is that information is provided about specific strengths and weaknesses, and progress can be documented. Administration time is 10-20 minutes per subtest and 50-100 minutes in total (Aiken & Groth-Marnat, 2006).

15. The *Brigance Inventory of Early Development – Second Edition* (Brigance IED-II, Brigance, 2004) is a norm- and criterion-referenced, English language instrument that assesses development in the areas of language, motor, academic-cognitive development, daily living, and social-emotional development. The division into subscales is not supported by factor analysis, which indicates a three-factor solution: understanding and expressing, movement and social activity, and academic-preacademic. Reliability coefficients range from .85 to .99. No information is given about the test floors and item gradients. It is possible to administer selected items, depending on the need for data. Data for many of the items can be completed by means of parent or teacher interviews (Early Childhood Measurement and Evaluation Resource Centre, 2007). This means that accommodations are not necessary, but also that the instrument is standardized to a limited extent. The usefulness of the results for intervention is enhanced by the fact that the instrument identifies specific strengths and weaknesses of the child. Administration time varies between 20 and 55 minutes (Early Childhood Measurement and Evaluation Resource Centre, 2007).

Two of the instruments found in the literature search were nonverbal.

11. The first is the *Snijders-Oomen Nonverbal Intelligence Test, Revised* (SON-R 2½-7). The instrument is available in the Dutch (Tellegen, Winkel, Wijnberg-Williams, & Laros, 1998a), English (Tellegen, Winkel, Wijnberg-Williams, & Laros, 1998b), German (Tellegen, Laros, & Petermann, 2007), French

(Tellegen, Laros, & Kiat, 2009), Czech (Tellegen, Laros, & Heider, 2008), Slovak (Tellegen, Laros, Kopcanova, Farkasova, & Dockal, 2009) and Portuguese (Laros, Tellegen, de Jesus, & Karino, 2011) languages. The subtests are mosaics, categories, puzzles, analogies, situations, and patterns. Although the SON-R 2½-7 generally has good psychometric qualities, the internal consistency reliability coefficients for the youngest age groups (2 years and 6 months and 3 years and 6 months) range from .41 to .81 for the subtests and from .68 to .90 for the composites. The test floor is 50, but the reliability in the low range of the norm score distribution is low for children between 2 years and 2 years and 6 months old (Tellegen et al., 1998a). The item gradients are adequate. The instrument is especially suitable for children with problems in verbal communication and language, and can also be applied for the developmental assessment of children without language problems. The built-in accommodations are instructions provided through demonstration and pointing. There are separate norms for deaf children. Although the instrument cannot be characterized as a learning-potential test, it does have some of the features of one: after every answer, feedback is given about the answer being right or wrong, and in case of a wrong answer, the right answer is shown to the child (Winkel, 1999). The response of the child to this feedback yields extra information that can be supportive in setting up an intervention plan. The administration time is 50-60 minutes (Tellegen et al., 1998a; 1998b).

16. The second nonverbal instrument is the *Leiter International Performance Scale – Revised* (Leiter-R; Roid & Miller, 1997). It is an English language instrument with nonverbal instructions and measures intelligence in the areas of reasoning, visualization, memory, and attention. This nonverbal instrument appears to be suitable for children with speech/language problems, hearing problems, or motor problems. The reliability coefficients for the subtests for 2-year-olds range from .71 to .94, with 2 out of 6 coefficients below .80. The reliability for the composites for 2- to 5-year-olds ranges from .88 to .92. The floors are inadequate for the majority of subtests for the 2-year-olds, but the item gradients are adequate. In the validity studies, the instrument was administered to children with speech delays; ADHD; English as a second language; and hearing, motor, or cognitive impairments. Information, in addition to the standard scores, can be obtained for low-functioning examinees by describing small incremental gains in skills over time using a scale based on a Rasch IRT model (Bradley-Johnson, 2001). The built-in accommodations are instructions given in mime, with the response by children through manipulating shapes, placing cards, or pointing. Only general instructions

for miming are given, thus allowing for a variability in administration that could affect the results. Items are not timed at age 2 years, which increases the suitability for young children with special needs. No special attention is paid to the supportiveness of test results for setting up an intervention plan. The administration time is 25-40 minutes (Bradley-Johnson, 2001).

The next two instruments were designed for use with children with motor impairments.

17. The Dutch version of the *Bayley Scales of Infant Development, Second Edition* (BSID-II; Van der Meulen, Ruiter, Spelberg, & Smrkovsky, 2002) has been adjusted for use with children with motor impairments aged 12 – 42 months (BSID-II-NL Low Motor; Ruiter, Nakken, Van der Meulen, & Lunenborg, 2010). The instrument has been evaluated positively by experts, although some materials appear to be too large for the youngest children. Pilot research shows that children with a motor impairment obtain, on average, a developmental index score of 5-10 points higher with the adjusted version as compared to the standard BSID-II, although there is no difference in observed scores between the two versions for the control group. This indicates that the adjustments have lowered the impact of motor impairments on the test results without changing the nature and difficulty of the instrument, thus supporting the applicability of the existing norms (Ruiter et al., 2010). Except for the adaptations, the instrument has the same characteristics as the BSID-II (Ruiter et al., 2010; Van der Meulen et al., 2002). The internal consistency reliability is low, the test floor is 55, the item gradients are inadequate, the supportiveness of the test scores as a basis for an intervention plan is limited, and administration time is 35-60 minutes.

18. The *Mayes Motor-Free Compilation* (MMFC; Mayes, 1999) is suitable for children with motor impairments and has a cognitive, social communication, and a language subscale. It is available in English. The administration is semi structured, and the instructions for the test administrator describe skills without elaborately describing what the administrator should do and say. The instrument has no norms, but the mental age is calculated using a formula. This is based on the base and ceiling of the administered items, the total number of items scored as passed, and the total number of items completed by the child. The interscorer reliability appears to be high, but other forms of reliability were not checked. Given the fact that there are no standardized scores, it is no surprise that there is no problem with the test floor and that nothing can be said about the item gradients. The instrument contains

built-in accommodations. An item is not counted as a failure when it is beyond the motor, vocal, or visual capability of a child with a physical impairment or if the child refuses to try an item. Instead, it is eliminated, and the child's score is prorated only on items appropriate for and attempted by the child (Mayes, 1999). Although one of the purposes of the MMFC is to assist in early intervention, there is no information about how this is to be done. An examination of the items makes clear that the instrument does not specifically yield information about the relative strengths of the child or about the child's learning potential. The administration time is 15-45 minutes.

In Table 2.2, the characteristics of the 18 instruments are summarized. The subtest's reliability across the full range of norm scores distribution appears to be problematic for 10 out of 18 instruments and moderate for a further three instruments. Internal consistency is good only for the DAYC. The reliability of the factors is generally better than the reliability of the subtests, as might be expected, but still problematic for three of the instruments for which the reliability was researched. Under test floor, the lowest possible index score is given, when available. If not, the judgment provided is based on the judgment found in reviews of the instrument, or "ni" (no information available) is indicated. According to Flanagan and Alfonso (1995), a test floor is appropriate if the lowest possible score is at least two standard deviations below the mean. Based on this criterion, the test floor is appropriate for nine of the instruments. For the other instruments, the floor was judged to be inadequate by the reviewer or no information was available. Only the SON-R-2½-7 and the Leiter-R have adequate item gradients, using the criterion as formulated by Bracken (1987) of no fewer than three raw score items per standard score standard deviation. For the other instruments, item gradients were inadequate or no information was available. With the exception of the instruments specifically designed for special groups and the ET 6-6, none of the instruments appears to be suitable for use with children with motor or visual impairments or describe standardized accommodations. Some instruments are specifically designed for use with children with motor impairments, but the psychometric properties of these instruments are not sufficiently researched. Some of the instruments are nonverbal or contain a nonverbal scale, such as the Stanford-Binet scales, the Kaufman-ABC-II, and the DAS-II/BAS-II. The idea of not counting items that are not suitable for a specific child, such as in the MMFC, is promising but needs to be described in more detail so it can be applied in a standardized way.

Table 2.2
Characteristics of the Instruments Relevant When Applied in the Assessment of Children With Special Needs

No. / Name	Subtest Reliability	Factor Reliability	Test Floor	Item Gradients	Accommodations Available	Information About Intervention	Duration (Minutes)
1. Bayley-III	-	±	55	-	Yes (not standardized)	No	50 – 90
2. Stanford-Binet 5th ed.	±	+	40	ni	Yes (not standardized)	No	15 – 75
3. WPPSI-III	-	-	55	-	Yes (short, not standardized)	No	35 – 42
4. Kaufman-ABC-II	-	+	+	ni	Nonverbal scale Teaching items	Yes	25
5. Woodcock-Johnson-III	-	±	-	-	Yes (not standardized)	Yes	45 – 90
6. BDI-2	-	±	40	-	Yes (not standardized)	No	60 – 90
7. Griffiths 0-2 / 12. Griffiths 2-8	ni	ni	+	ni	No	No	50 – 60
8. BAS-II / 9. DAS-II	-	-	+	ni	Nonverbal score Individualization possible	Yes	20 – 39
10. PEP-3	±	+	ni	ni	Suitable for children with ASD	Yes	45 – 90

(Table continues)

Table 2.2 ***(continued)***

No. / Name	Subtest reliability	Factor reliability	Test floor	Item gradients	Accommodations available	Information about intervention	Duration (minutes)
11. SON-R 2½-7	-	-	50	+	Nonverbal instrument	Yes	50 – 60
13. ET 6-6	ni	ni	ni	ni	Suitable for children with motor impairment	No	45
14. DAYC	+	na	-	ni	No	Yes	50 – 100
15. Brigance IED-II	±	na	ni	ni	Not necessary (interview option)	Yes	20 – 55
16. Leiter-R	-	±	-	+	Nonverbal instrument	No	25 – 40
17. BSID-II-NL-Low Motor	-	na	55	-	Suitable for children with motor impairment	No	35 – 60
18. MMFC	ni	ni	na	ni	Suitable for children with motor impairment	No	15 – 45

Note. - = low / inadequate; ± = moderate; + = high / adequate; ni = no information available; na = not applicable.

The amount of information obtained that is supportive in setting up an intervention plan is limited for instruments that generate a standardized score only. Some instruments do generate extra information, however. The Kaufman-ABC-II, for example, gives a child examples and a second attempt if an item appears too difficult. The subsequent response of the child provides information about the learning potential. Other instruments that explicitly yield information that is expected to be useful when setting up interventions are the Woodcock-Johnson III, the DAS-II/BAS-II, the PEP-3, the DAYC, the Brigance IED-II, the SON-R-2½-7, and the MMFC.

As can be seen in Table 2.2, the *administration time* varies widely among the instruments. Most instruments have an administration time of less than 60 minutes, but for some instruments the duration is a problem, such as the Bayley-III, Woodcock- Johnson-III, BDI-2, PEP-3, and the DAYC.

2.4 Discussion

The aim of this article was to offer an overview of contemporary instruments for the developmental assessment of children younger than the age of 4 years and their applicability to children with different kinds of special needs. The results show that for children younger than 4 years of age, a variety of standardized instruments are available, and the characteristics of these instruments vary substantially in terms of reliability, test floors, item gradients, applicability for functional impairments, usefulness for setting up an intervention, and administration time. The Stanford-Binet and the SON-R-2½-7 appear to have relatively good qualities in these areas.

It can be concluded that no test is suitable for all different types of special needs. For children with a hearing or speech/language impairment, there are two nonverbal instruments: the SON-R-2½-7 and the Leiter-R. In addition, the Kaufman-ABC-II, and the BAS-II and DAS-II are suitable, because they contain a nonverbal score. All these instruments have a starting age of 2 years or older, which implies that there are no suitable instruments for the younger children who need a nonverbal instrument. However, other instruments without a nonverbal score and with a starting age younger than the age of 2 years, such as the Bayley-III, BDI-2, and Griffiths 0-2, might be suitable. They do not contain a lot of language because all children younger than the age of 2 years have limited language comprehension and production.

Two instruments are specifically designed for children with motor impairments (the BSID-II-NL Low motor and the MMFC). Additionally, the ET 6-6 appears to be suitable for this target group. However, all three instruments have limitations in terms of reliability, or their reliability has not yet been researched. No instrument was found in the literature search that is designed for children with visual impairments. Our conclusion is that there is no suitable instrument available for this group of children.

It very much depends on the characteristics of the specific child being tested whether an appropriate instrument is available. For example, if the reliability of a test is appropriate except for very young children, the test is suitable for older children. The same goes for item gradients. As long as there are no item gradient violations in the age range of the child to be tested, there is no problem. A high test floor is only inappropriate when the instrument is being used for testing a child scoring significantly below average. It also should be noted that an instrument without accommodations is only inappropriate for children who have a functional impairment.

Our conclusion is that a range of instruments exists for the developmental assessment of children younger than the age of 4 years, all of which have their strengths and weaknesses. Work needs to be done to improve the qualities of the instruments available for children younger than 4 years of age in general, and for children with a motor impairment and children younger than 2 years of age in particular. Furthermore, for those children with a visual impairment, no suitable instrument yet exists.

2.4.1 Limitations

A few of the limitations of the current review should be kept in mind. First, we only included those instruments that were found through the literature search. It is possible that more instruments exist that fit within the scope of this article but that no test reviews or articles that reported on research into the use of these instruments have been published in peer-reviewed journals in English from 1995 onwards. There is a good chance that relevant instruments exist that were not included in this review as a consequence of the methods used. However, including all the existing developmental assessment instruments would have been impossible because of the wide range of instruments available. Furthermore, our search was intended to provide an overview of only high-quality and widely used instruments, and a published article in an international peer-reviewed journal is an indication of high-quality research. Nevertheless, it is important to bear in mind that this is not a comprehensive overview of standardized developmental assessment instruments.

Second, the information about the instruments in this review is based on information available in the (online) library, including test reviews on the website of test publishers and, in some cases, the manuals of the instruments. Although these sources give quite a thorough description of the instruments, it is possible that some information is missing as a consequence of not having been able to read and make use of the test manual or other relevant information about the instrument. In addition, when the instrument was designed for a broader age range, specific information on the various issues of interest to us was not always available specifically for young children.

Third, as we already emphasized before, this review is not an extensive overview of the psychometric properties of the instruments. If this had been the aim, we would have discussed other characteristics as well, such as the validity of the instruments and the different kinds of reliability. Instead, this review offers an evaluation of widely used standardized instruments in relation to frequently mentioned issues in the developmental assessment of young children with special needs.

Finally, the evaluation of the usefulness of test results for setting up intervention strategies should be interpreted with caution. Any such evaluation is very complex and actually would require an in-depth review of all the instruments described and their results. The information provided in this review is intended only to provide a basic idea about the nature of the test results and their potential usefulness for the purposes of setting up a plan of intervention.

2.4.2 Future Research

Future research is necessary to develop an appropriate set of instruments for young children with special needs. The psychometric properties of instruments that have not yet been thoroughly and sufficiently examined in this regard need to be studied.

More specifically, it appears that instruments available for the developmental assessment of children with motor impairments have insufficient reliability or have not yet undergone research in this regard. Furthermore, any research should be done with the aim in mind of developing a suitable instrument for children with visual impairments. The suitability of existing instruments for use with different clinical groups needs to be studied more carefully than has been done so far, taking into account the large differences that exist among children with special needs.

The usefulness of test results when setting up an intervention plan should be researched empirically. In this article, their usefulness was judged simply based on the nature of the test results. However, the presence of information about, for example, a child's response to feedback does not entail that this information will be, by definition, supportive in intervention planning. Planning also depends on the qualities of the specialist as well as the characteristics of the assessment process of which the administration of a standardized instrument is just one part.

All future research on developmental assessment instruments and especially on the use of test results in intervention planning should be expected to show whether assessment can indeed be used for different purposes simultaneously. If this appears not to be possible, assessment to obtain information for intervention purposes should be done separately from assessment for eligibility purposes, for example. Instruments would then need to be developed for these specific purposes. As long as sufficient research has not been done, however, it is preferable to combine all the assessment purposes for reasons of efficiency.

Chapter 3

Accommodating the Bayley-III with regard to motor and/or visual impairment: A comparative pilot study*

* This chapter has been accepted for publication as: Visser, L., Ruiter, S. A. J., Van der Meulen, B. F., Ruijssenaars, A. J. J. M., & Timmerman, M. E. (in press). Accommodating the Bayley-III with regard to motor and/or visual impairment: A comparative pilot study. *Pediatric Physical Therapy.*

Abstract

Purpose: This study aimed at assessing the validity and usefulness of the Bayley-III Low Motor/Vision accommodated version. Accommodations are adaptations to minimize impairment bias, without altering what the test measures. Of the items, 66% have Low motor accommodations like enlarged materials; 62% have Low vision accommodations. *Method:* Using a within-subject design, we tested 19 children with the accommodated and standard Bayley-III, in randomly counterbalanced order. The children had motor and/or visual impairment and a calendar age between 22 and 90 months. The test administrators completed an evaluation form. *Results:* A subgroup of children benefitted from the accommodations; 2 children obtained a large raw score difference. Test administrators considered the accommodations as practicable, and advantageous for a majority of children. *Conclusion:* The Low Motor/Vision accommodated version seems to validly assess the development of this target population. Future, larger-scale research should study whether the accommodations improve the construct validity of the Bayley-III.

3.1 Introduction

Estimates of the prevalence of special needs in young children (0-3 years) vary, generally ranging from 5% to 10% of the population in the United States and the Netherlands (Broer van Dijk - Van der Hulst et al., 2005; Rydz, Shevell, Majnemer, & Oskoui, 2005). Professionals use standardized instruments to objectively assess the development of children with special needs. This is in conformity with national regulations in, for example, Europe and the United States (Individuals with Disabilities Education Improvement Act, 2004a; Parveva, De Coster, & Noorani, 2009). Substantial numbers of the children with special needs have a motor and/or visual impairment (Wever, De Klerk, & Van der Loos, 2006). It is essential that appropriate and fair instruments are available for this group (Bradley-Johnson, 1994; M. R. Johnson, Wilhelm, Eisert, & Halperin-Phillips, 2001; Miller & Skillman, 2003). This group is especially in need of developmental assessment, and test results often have a large influence on choices regarding care and education.

However, many professionals indicate that suitable instruments are lacking (Groenveld, 1990; Neisworth & Bagnato, 2004; Skovgaard, Houmann, Landorph, & Christiansen, 2004; Snow & Van Hemel, 2008; Visser, Ruiter, Van der Meulen, Ruijssenaars, & Timmerman, 2012). Applying the standard procedures when testing children with a motor and/or visual impairment seriously threatens the validity of the test results. Most instruments that measure cognitive development in children rely heavily on motor skills, especially in the case of young children, whose language skills are not yet well developed (Ruiter, Nakken, Van der Meulen, & Lunenborg, 2010). Test manuals often provide suggestions for adaptations, but using unstandardized adaptations may introduce additional sources of measurement error and bias, and therefore preclude interpreting the test results using the standard norms (American Educational Research Association [AERA], American Psychological Association [APA], & National Council on Measurement in Education [NCME], 1999).

To meet the need for appropriate instruments for children with impairments, one could develop a new instrument for a population of children with a specific impairment. This approach has been taken, for example, in the Mayes Motor-Free Compilation (MMFC; Mayes, 1999) for children with motor impairments. Alternatively, one may accommodate an existing, well-developed and high quality instrument that has been designed for the entire population of young children. Accommodating an instrument implies that changes are made to the format, response possibilities, test circumstances, and/or procedures in order to minimize impairment bias, without altering what it measures (Batshaw Claire, Church, & Batshaw, 2007; Thurlow, Elliot, & Ysseldyke, 2003). In other words, accommodations do not change the content and difficulty of the test items, but they do increase the construct validity by decreasing the influence of an impairment on the test results. Studies are needed to assess the impact that accommodations have on test validity. If changes to a test are indeed just accommodations, it will then not be necessary to conduct large-scale and time-consuming standardization research for a specific group of children. The original norm tables will apply, hence allowing for a direct comparison of the test results of children with an impairment with the results of typical children of the same calendar age.

In the current study, accommodations were made to the Bayley Scales of Infant and Toddler Development, Third Edition (Bayley-III; Bayley, 2006a) to increase its suitability for assessing children with a motor and/or visual impairment. The aim of the resulting Low Motor/Vision accommodated version is to enhance children's prospects of being able to show their cognitive, language, and motor skills in a test situation. The term "Low" refers to the amount of motor and visual components in the items. We removed the motor and visual components as much as possible in order to obtain an accommodated version. For example, the motor component (e.g., pointing) was eliminated in items designed to measure cognitive ability (e.g., connecting similar pictures). Since our intention was not to change the item content and difficulty, we will be using the term "accommodations" to describe the changes made to the test. The result should be that children for whom the standard version is suitable have equal scores on the accommodated and standard versions of the item (apart from measurement error). We expected that the construct validity of the resulting measurement would increase as a result of a more precise estimation of the competencies of interest. If this proved to be the case, then the use of the standardized Low Motor/Vision accommodated version, combined with the original norm tables, should enable professionals to compare the development of a child with a motor and/or visual impairment with the typical development of children with the same calendar age.

Comparable research has been done with the Dutch second version of the Bayley Scales of Infant Development (BSID-II-NL; Van der Meulen, Ruiter, Spelberg, & Smrkovsky, 2002). Pilot research into this Low Motor and Low Vision version suggests that the accommodations make the test easier to administer, more engaging for the children, and produce more valid outcomes (Ruiter et al., 2010; Ruiter, Nakken, Janssen, Van der Meulen, & Looijestijn, 2011).

The purpose of the present study was to evaluate whether the Low Motor/Vision accommodated version of the Bayley-III would yield more valid test results, when testing children with a motor and/or visual impairment, than the standard version of the instrument. Furthermore, we studied whether the instrument was practicable for the person administering the test.

3.2 Method

3.2.1 Study Design

We evaluated the Low Motor/Vision accommodated version of the Bayley-III in a pilot study using a within-subject design. We tested the children once with the Low Motor/Vision accommodated version and once with the standard version of the Bayley-III. The average time interval was two weeks (range 3 to 22 days, with two outliers of 28 days for child 2 and 45 days for child 8). The target interval was 7 to 14 days, but for organizational reasons it proved to be impossible to meet this target for all the children. However, the impact of this variation in interval length would appear to be limited: the impaired development of the children in combination with their relatively older calendar age (i.e., 22 months or older) should result in no great difference in developmental level being expected within a one-month period.

We also counterbalanced the order in which the children were tested. Eleven children were first tested with the Low Motor/Vision accommodated version and then with the standard version; eight children were tested in the reverse order. As a consequence of age-specific starting points, and of reversal and discontinue rules in the Bayley-III, only part of the items per scale were administered. Note that the actual items administered to a child could differ across the two test administrations as a consequence of differences in responses to the test items.

The referring developmental psychologist filled in a short referral form for each child. A test administrator then tested the child. The nine test administrators in our study were advanced university students in special needs education or psychology, who had gone through an intensive training session to learn how to administer and score the test. After this training session, the test administrators conducted a practice test with five children before starting to test for our research data. Two of these five test administrations were observed via video recording by one of the two principal researchers, who are professionals in administering and training for administering the Bayley-III. For each video, the researcher offered feedback about the interaction with the child (e.g., how to deal with shyness), the way of administering the test items (e.g., “You should remove the colored disks from the picture after each answer by the child”), and the scoring (e.g., “I saw that you also administered item number X, but the stopping rule should already have come into effect at that point”). No serious errors were observed for any of the administrators, and the feedback was limited to only a few feedback points. During the entire testing period, the principal investigators and the test administrators held regular meetings. In those meetings, questions were asked and experiences shared, including discussions about certain items that appeared to be difficult to score in some cases. The test administrator was the same person across test sessions for 10 of the children and was different for nine of the others. The tests took place in the Netherlands at a rehabilitation center or an organization supporting persons with a visual impairment, which the children attended multiple days a week. A parent or teacher who knew the child well was present during the test.

3.2.2 Participants

Nineteen children participated in this study. The children were referred by the developmental psychologist of the referring organization. The first inclusion criterion for children participating in the study was a diagnosis of mild to severe motor impairment affecting arm and/or hand movement, and/or a diagnosed or suspected visual impairment. Note that a child with a motor impairment affecting only a lower extremity does not meet the inclusion criterion. We expect that such impairments would have no effect on the test score in the standard version of the test, and therefore the Low Motor accommodations only relate to the hands and arms, not the legs. Visual impairment was defined broadly, including disorders of the eye as well as visual impairment due to damage to the brain (e.g., cerebral visual impairment). The developmental psychologist provided the information about diagnoses and impairment via the referral form. We did not obtain any information about the process leading to the diagnosis such as who had made the diagnosis and which instruments had been used.

Additional inclusion criteria were: (a) calendar age between 6 months and 10 years; (b) presumed developmental age between 1 and 42 months (age range Bayley-III); (c) ability of the child to sit upright in a chair or wheelchair so that a table could be used to work upon; (d) ability of the child to use at least one hand; and (e) some visual perception ability (hence blind children were excluded). The last three criteria describe the minimum abilities needed to perform the actions required for the test items.

Five different organizations referred children on the basis of the inclusion criteria. All the referred children were tested. The test results were used simultaneously for our research and in the diagnostic process performed by that organization. One child was tested with the Low Motor/Vision accommodated version, but could not be tested with the standard version as a consequence of moving out of the region. We excluded this child's data from the study, and the child was not included in the total number of 19.

The mean calendar age of the children at the first testing session was 38 months (range 22-90 months), and there were 11 boys and 8 girls. Table 3.1 shows detailed information, provided by the developmental psychologist, about the children in terms of calendar age, gender, type of referral organization, diagnoses, and impairment. The children numbered one to eight in Table 3.1 had a motor impairment (n = 8); the children numbered nine to 19 had a motor and visual impairment (n = 11). We divided the information about the impairment into three categories: disorder or disease (based on the International Classification of Diseases – 10th edition, World Health Organization [WHO], 1992), body functions and structures, and activities (both based on the International Classification of Functioning, Disability and Health – Children and Youth version, WHO, 2007). A "-" means that the child was not diagnosed with any specific disease or disorder, or that the referral form did not specify any information about the implications of the impairment for the activities of the child. In all of the cases, the referring organization had classified the child as having a motor or visual impairment, and thus granting access to their services.

Both the standard Bayley-III and the Low Motor/Vision accommodated version were administered to all children. Both versions consist of five scales. The children with a primary visual impairment were administered all five scales. The children with a primary motor impairment were not administered the two scales pertaining to motor abilities. The Motor scales were not accommodated for any motor impairment, because that would have threatened the construct validity of these scales. The impaired skill, in this case, is meant to be measured.

For some of the children (n = 6) the test could not be carried out completely due to time constraints of the organization involved and tiredness of the child, resulting in an early completion of the test for these children. Table 3.1 shows which scales were administered per child.

Table 3.1

Characteristics of the n = 19 Children in the Pilot Sample. Children Numbered 1 - 8 have a Motor Impairment; Children Numbered 9 - 19 have a Motor and Visual Impairment.

Child number	Calendar age (m;d)	Gender	Referring organization	Disorder or disease [a]	Body functions and structures [b]	Activities [b]	Administered subscales Bayley-III
1	26;15	girl	Rehab.	Neonatal convulsions	Developmental delay Motor impairment Disorder of tonus regulation	Prefers to use left hand	Cog, RC
2	27;0	girl	Rehab.	Perinatal asphyxia Ischemic brain damage Cerebral palsy	Bilateral spastic cerebral palsy, GMFCS 3	Able to walk	Cog, RC, EC
3	27;4	boy	Rehab.	Perinatal porencephalic cyst, left frontal	Developmental delay Motor impairment in right upper extremity Hypokinesia Hypotonia of the torso	Right hand in fist, child uses this hand sometimes	Cog, RC, EC
4	27;25	boy	Rehab.	-	Psychomotor developmental delay Slow processing of stimuli	-	Cog
5	33;23	girl	Rehab.	Premature birth with bleeding in ventricular system Cerebral palsy	Bilateral cerebral palsy, especially legs are affected	Walks with walker	Cog, RC, EC
6	41;11	boy	Rehab.	Cerebral palsy	Spastic bilateral cerebral palsy, GMFCS 4	Can play with two hands when in good form Cooperation between the two hands is tiring and not smooth Movement jerky with grasping and letting go	Cog, RC, EC

(Table continues)

Table 3.1 (Continued)

Child number	Calendar age (m;d)	Gender	Referring organization	Disorder or disease [a]	Body functions and structures [b]	Activities [b]	Administered subscales Bayley-III
7	47;13	boy	Rehab.	Cerebral palsy	Spastic bilateral cerebral palsy, GMFCS 4	Impaired torso balance affecting alertness Uses both hands Often uses palmar grasp; decreased force and coordination when using more advanced grasping	Cog, RC, EC
8	90;6	boy	ID	Epilepsy	Developmental delay Hypotonia Short attention span	Difficulty sitting for extended period of time	Cog, RC, EC
9	22;19	boy	Rehab.	Cerebral palsy	Unilateral spastic cerebral palsy Increased tonus / spasticity at the right side Minor visual impairment	Right hand often in fist, rarely used Sits with support in adjusted chair	Cog, RC, EC
10	29;27	girl	Rehab.	Cerebral visual impairment	Problems with visual information processing Psychomotor developmental delay	Needs time to respond to stimuli	Cog, RC, EC
11	30;26	girl	Rehab.	IFAP syndrome Epilepsy	Severe developmental delay Motor impairment Visual impairment	Does not walk or crawl Uses glasses	Cog
12	31;28	girl	Rehab.	-	Delayed motor development Hypotonia Visual acuity 0.02 – 0.08	Optimal visual capacity in faint light	Cog, RC, EC
13	32;13	boy	Rehab.	Mowat-Wilson syndrome Absence of neurohypophysis Possible optic nerve hypoplasia at right side	Motor impairment Visual impairment	Needs time to grasp Difficulty with visual fixation	Cog, RC, EC

(Table continues)

Table 3.1 (Continued)

Child number	Calendar age (m;d)	Gender	Referring organization	Disorder or disease [a]	Body functions and structures [b]	Activities [b]	Administered subscales Bayley-III
14	34;23	boy	Rehab.	-	Developmental delay Motor impairment Visual impairment	Sensitive to stimuli Has glasses but does not tolerate them Able to walk	Cog
15	35;26	boy	Rehab.	Unknown syndrome Palatoschisis	Psychomotor retardation Mild impairment in vision and hearing	Unable to move from place to place independently	Cog, EC
16	36;5	girl	Rehab.	Hydrocephalus Microcephaly Epilepsy	Psychomotor developmental delay Visual impairment	-	Cog, RC
17	39;7	boy	Rehab.	West syndrome	Developmental delay Impaired registration of taste, hearing and visual stimuli Motor impairment Visual impairment Short attention span	Finds it difficult to stay seated Needs clear instructions	Cog, RC, EC
18	53;16	boy	Visual imp.	Infantile encephalopathy Polymicrogyria Velo-Cardio-Facial Syndrome Possible Cerebral Visual Impairment	Severe psychomotor retardation Spasticity Hypotonia Auditory and visual information processing problems with a normal visual acuity Fine motor skills moderately developed Gross motor impairment	Looks at objects while playing, but does not look during social interaction Able to crawl Not able to walk	Cog, RC, EC, FM, GM
19	56;12	girl	Visual imp.	Microcephaly Palatoschisis Epilepsy Possible Cerebral Visual Impairment	Severe psychomotor retardation Hypotonia Flat feet Epileptic seizures with severe shaking Visual acuity 0.20-0.25 with glasses	Does not crawl, stand or walk Able to play seated Easily distracted Slow processing of sensory stimuli	Cog, RC, EC, FM, GM

Note.(m;d): (months; days); Rehab.: Rehabilitation Centre; ID: Organization supporting people with Intellectual Disabilities; Visual imp.: Organization supporting people with Visual impairment; GMFCS: Gross Motor Function Classification System; Cog: Cognition scale; RC: Receptive Communication scale; EC: Expressive Communication scale; FM: Fine Motor scale; GM: Gross Motor scale.

[a]In line with the International Classification of Diseases, ICD-10 (WHO, 1992).

[b]In line with the International Classification of Functioning, Disability and Health for Children and Youth, ICF-CY (WHO, 2007).

3.2.3 Instruments

The standard version of the Bayley-III is an individually administered instrument that assesses the psychological and psychomotor development of children with a developmental age of between 1 and 42 months. The instrument consists of the scales of Cognition, Receptive Communication, Expressive Communication, Fine Motor Development, and Gross Motor Development. Items are scored positively (1) when a child has shown the target behavior and negatively (0) when not. The starting point depends on the calendar age of the child, and the highest starting point is used when a child is more than 42 months of calendar age. Items before the starting point are then not administered and are automatically scored as 1. The stopping rule is to stop after five consecutive items have been scored 0, and all items after the final administered item are not administered and are automatically scored as 0. The domains of social-emotional development and adaptive behavior were assessable on the basis of primary caregiver responses to a questionnaire, which was not part of the current research. In this study we used the experimental version of the Dutch Bayley-III, which is identical to the American version, except for the language. Standardization research in the Netherlands is currently ongoing.

The standardization sample of the Bayley-III in the United States included 1,700 children. Validity data were given in the form of moderate to high correlations of Bayley-III test scores with scores on other instruments. The internal consistency and test-retest stability appeared to be good (Bayley, 2006b).

Bayley-III Low Motor/Vision accommodated version is similar to the standard version of the test except for the accommodations made to test procedures, item instructions, and play materials. The scoring procedure is also identical to that of the standard version. The accommodations were based on those of the Low Motor and Low Vision accommodated versions of the Dutch Second Edition of the Bayley Scales of Infant Development (Ruiter et al., 2010; Ruiter et al., 2011), complemented with accommodations for the new Bayley-III items, which were developed in close cooperation with developmental psychologists working in the field. We did not delete any items.

The Low Motor accommodations were made for the full age range of the Cognition and Language scales. The Low Vision accommodations were made for the full age range of all five scales. We were able to combine the Low Motor and Low Vision accommodations into one test version, which then had the clear advantage that the test was suitable for children with both motor and vision problems.

Items were accommodated in terms of materials, item instructions, or both. Table 3.2 gives the number of items that were accommodated and the total number of items per scale of the Bayley-III. If possible, we made larger versions of standard test materials that were too small for a child with a motor impairment to handle because of the need for using mature fine motor skills. We added a placemat colored dark blue and changed the color of most materials to yellow, which provides optimal color contrast with the dark blue placemat.

Table 3.2
Number of Items that were (not) Accommodated in the Bayley-III Low Motor/Vision, per Type of Accommodation

	Cognition		Receptive Communication		Expressive Communication		Fine Motor Development	Gross Motor Development
	LM	LVi	LM	LVi	LM	LVi	LVi	LVi
Materials only	15	29	3	27	16	14	7	14
Instructions only	24	15	4	2	1	0	28	13
Materials & Instructions	30	22	31	4	0	2	26	0
None	22	25	11	16	31	32	5	45
Total	91	91	49	49	48	48	66	72

Note. LM: Low Motor accommodation; LVi: Low Vision accommodation.

There were three categories of accommodations to the item instructions: (1) the use of eye pointing instead of finger pointing (Low Motor); (2) support of the child's elbow by the test administrator (Low Motor); and (3) placing objects and pictures closer to the child, if necessary (Low Vision). We applied these accommodations to each applicable item in the Cognition and Communication scales (both Low Motor and Low Vision) and Motor scales (Low Vision only).

In addition to the accommodations in materials and instructions, we accommodated the test procedure by eliminating the time limits for all items, because a motor and/or visual impairment commonly results in more time needed to complete a task. Accommodations to the test procedure thus also apply to those items without any accommodations to the materials or instructions.

Evaluation form. The test administrator filled in an evaluation form to determine whether the accommodations were practicable for the person administering the test and suitable for the specific child being tested. We defined practicable as "able to be put into practice successfully" (Oxford Dictionaries Online, 2012a) and suitable as "right or appropriate for a particular person, purpose, or situation" (Oxford Dictionaries Online, 2012b), in this case for the assessment of a child with a motor and/or visual impairment. If the test administrator differed across test sessions, the form was completed by the person who administered the Low Motor/Vision accommodated version. If a developmental psychologist or teacher observed the test administration, their feedback was included. The questions in the evaluation form were: “Do the test results from the accommodated version correspond with your view of the developmental level of this specific child?”, “Were the Low Motor/Vision accommodations practicable when testing this child?”, and “What were the advantages of the Low Motor/Vision accommodations for this child when compared to the standard version?”. We also asked for additional comments, and we asked whether the test manual and item instructions were clear and unambiguous, and whether the record form contained all necessary information.

3.2.4 Analysis

We took into consideration the *raw score difference* per scale, which is computed as the raw score on the Low Motor/Vision accommodated version minus the raw score on the standard version. Hence, a positive figure indicates a higher score on the Low Motor/Vision accommodated version than on the standard version. The total raw score was calculated following the default scoring rules of the Bayley-III.

Noting that the total raw score also included non-accommodated items, we also took into consideration the *percentage score difference on adjusted items*. Adjusted items are accommodated items that were actually administered to that specific child using both versions. The percentage indicates how large the improvement (or decline) in test score is, in relation to the total number of adjusted items. In identifying the adjusted items per child, we took into account the impairment of the child. Thus, for children with a motor impairment, we only took into consideration the items with a Low Motor accommodation. For children with a motor as well as a visual impairment, we took into consideration the items with a Low Motor and/or Low Vision accommodation.

We used the percentage score difference rather than the absolute difference, because a score difference of, for example, 3 is a large difference, when only 9 accommodated items are administered, but not so large when 25 items are administered. The reason for considering both the *raw score difference* and the *percentage score difference on adjusted items* as outcome measurements is that both are clinically relevant. The *percentage score difference on adjusted items* is a very clean measurement of the influence of the accommodations. The *raw score* difference is relevant because the raw score is used in daily practice as a basis for the test results. If a child is able to complete an item as a consequence of accommodations, this may influence the course of the test administration. If the discontinue rule is not met at the same point that it would be in the standard version, the child gets the chance to show his or her abilities on items higher on the scale. The *raw score difference* can therefore be larger than the *score difference on adjusted items*.

We expected a higher score on the Low Motor/Vision accommodated version when compared to the standard version, because this would indicate that the child benefited from the accommodations. We used the one-sided one-sample Wilcoxon Signed Rank test adopting a significance level of 0.05 to test whether the median of the *raw score difference* and the median of the *percentage score difference on adjusted items* were significantly larger than zero. With this test, we examined whether support is found for the hypothesis that scores on the Low Motor/Vision accommodated version would be larger than the scores on the standard version in the target population of children.

To answer the research questions of whether the instrument was suitable for the children and practicable for the test administrator, we summarized the answers to the questions in the evaluation form. We identified areas of improvement on the basis of the results of this study.

3.3 Results

3.3.1 Test Results

We have summarized the test results in Table 3.3. This table shows the total raw scores on the standard version and the Low Motor/Vision accommodated version, the *raw score difference* (Raw score diff.) and the *percentage score difference on adjusted items* (% score diff. adj. items), per child and per subscale.

From the Wilcoxon Signed Rank tests on the scales of Cognition, and Receptive and Expressive Communication, it appeared that the median was not significantly larger than zero ($p = 0.432$, $p = 0.224$, and $p = 0.340$, respectively). We did the same test on the *percentage score difference on adjusted items* and these results were also not significant ($p = 0.101$, $p = 0.378$, $p = 0.104$, respectively).

As can be seen in Table 3.3, for all three scales the *raw score difference* indicates that some children obtained equal scores on both versions, some children obtained a higher score on the Low Motor/Vision accommodated version, and some children obtained a higher score on the standard version of the Bayley-III. The two children with a large *raw score difference* on the Cognition scale in favor of the Low Motor/Vision accommodated version (child 2 and child 4) both have a motor impairment and no visual impairment. The test reports revealed that both children had cooperated well during both test administrations. Alertness of the child as a confounding factor had thus probably not played a large role in the test results. The reports also revealed that child 2 had clearly benefitted from the enlarged materials. Except for the motor impairment, there is not much overlap in type of impairment: child 2 has cerebral palsy, while child 4 has psychomotor developmental delay (see also Table 3.1). Although the children were both 27 months old at the time of testing, the raw scores are not in the same range, which means that these two children were largely not assessed the same range of items. Therefore, it cannot be deducted from the current data why some children do have a high score difference. The same is valid for child 8 and child 9, who both obtained a relatively large raw score difference on the Expressive Communication scale: no overlap in specific impairment or assessed items can be found for these children.

The *percentage score difference on adjusted items* shows that on the Cognition and Expressive Communication scales, some children obtain a substantially higher score on adjusted items in the Low Motor/Vision accommodated version than on the standard version (with 9 and 5 children, respectively, showing an increase). This large benefit for some children is reflected in average percentages of change in the scores on adjusted items of 6% and 11% for Cognition and Expressive Communication, respectively.

The Motor scales were administered to child 18 and child 19 only; both had a visual impairment. We did not include these test results on the Motor scales in Table 3.3. Child 18 was administered the Low Motor/Vision accommodated version first and scored higher on the standard version for the Fine Motor scale: the *raw score difference* was -5, and the *percentage score difference on adjusted items* was -22% (-4/18). Child 19 was administered the standard version first and scored higher on the Low Motor/Vision accommodated version for the Fine Motor scale: the *raw score difference* was 2, and the *percentage score difference on adjusted items* was 22% (2/9). For the Gross Motor scale, Child 18 had a *raw score difference* of 3 and a *percentage score difference on adjusted items* of 50% (3/6). Child 19 had a *raw score difference* of -4 and a *percentage score difference on adjusted items* of -33% (-3/9), scoring lower on the Low Motor/Vision accommodated version.

Table 3.3
Test Results on the Standard Version and the Low Motor/Vision Accommodated Version

Child	Cognition			Receptive Communication			Expressive Communication		
Number (Impairment)	Raw score LM/LVi – Stand.	Raw score diff. [a]	% score diff. adj. items [b]	Raw score LM/LVi – Stand.	Raw score diff.	% score diff. adj. items	Raw score LM/LVi – Stand.	Raw score diff.	% score diff. adj. items
1 (M) [1]	36 – 34	2	20% (2/10)	11 – 11	0	0% (0/7)			
2 (M) [1]	49 – 38	11	43% (6/14)	24 – 25	-1	-9% (-1/11)	22 – 22	0	67% (2/3)
3 (M) [2]	28 – 27	1	6% (1/18)	13 – 13	0	0% (0/8)	13 – 13	0	0% (0/1)
4 (M) [2]	32 – 22	10	25% (4/16)						
5 (M) [2]	71– 68	3	18% (2/11)	36 – 34	2	9% (2/23)	45 – 46	-1	-8% (-1/12)
6 (M) [2]	72 – 76	-4	-22% (-2/9)	39 – 39	0	0% (0/17)	42 – 42	0	7% (1/14)
7 (M) [2]	65– 67	-2	-8% (-1/12)	32 – 31	1	7% (1/15)	35 – 33	2	11% (1/9)
8 (M) [1]	66 – 67	-1	-7% (-1/15)	31 – 27	4	19% (4/21)	31 – 25	6	57% (4/7)
9 (MV) [1]	41 – 39	2	21% (3/14)	12 – 11	1	14% (1/7)	15 – 9	6	33% (1/3)
10 (MV) [2]	50 – 52	-2	-8% (-2/25)	22 – 25	-3	-21% (-3/14)	23 – 25	-2	0% (0/4)
11 (MV) [1]	15 – 22	-7	7% (1/14)						
12 (MV) [2]	67 – 68	-1	0% (0/13)	36 – 39	-3	-12% (-3/26)	35 – 40	-5	-10% (-1/10)
13 (MV) [2]	30 – 30	0	0% (0/18)	9 – 10	-1	0% (0/4)	11 – 11	0	0% (0/1)
14 (MV) [1]	48 – 50	-2	-10% (-2/20)						
15 (MV) [2]	57 – 54	3	14% (2/14)				16 – 15	1	0% (0/3)
16 (MV) [1]	57 – 57	0	0% (0/23)	25 – 22	3	25% (2/8)			
17 (MV) [2]	58 – 61	-3	-12% (-3/26)	31 – 32	-1	-11% (-1/9)	33 – 34	-1	-10% (-1/10)
18 (MV) [2]	30 – 30	0	0% (0/24)	8 – 7	1	25% (1/4)	3 – 4	-1	0% (0/0)
19 (MV) [1]	24 – 24	0	20% (2/10)	7 – 4	3	-25% (-1/4)	4 – 3	1	0% (0/0)
Average		0.5	6%		0.4	1%		0.4	11%

Note. M: Motor; MV: Motor and Visual; LM/LVi: Low Motor/Vision accommodated version; Stand.: Standard version ; diff.: difference.
[1]Standard version administered first. [2]LM/LVi version administered first.
[a] A score difference is calculated by subtracting the score on the standard version from the score on the Low Motor/Vision version. E.g., child 6 obtained a raw score of 72 on the Cognition scale for the Low Motor/Vision version and a raw score of 76 on the standard version. The raw score difference is therefore 72 - 76 = -4.Empty cells indicate that the scale concerned was not administered to that child.
[b] % score diff. adj. items: percentage score difference on adjusted items. This percentage indicates how large the improvement (or decline) in test score is, in relation to the total number of adjusted items. Adjusted items are accommodated items that were actually administered to that specific child using both versions.

The Bayley-III provides age equivalents for each raw score. An age equivalent indicates the average of the ages (in months) at which children in the population obtain that particular raw score. We calculated the age equivalent differences between the two versions of the test (i.e., as Low Motor/Vision minus standard). For the Cognition scale, the range of age equivalent differences in the current sample was -4 months (i.e., standard version age equivalent 4 months higher than the Low Motor/Vision version) to 5 months (i.e., Low Motor/Vision version age equivalent 5 months higher than the standard version). For the Receptive Communication scale, this range was -7 months to 5 months; for Expressive Communication -9 months to 5 months. Thus, for the Cognition, Receptive Communication and Expressive Communication scales, developmental age equivalents of the accommodated versions were up 5 months higher than the one belonging to the standard version, which implies a clinically significant difference. This result should be interpreted with caution, however, because it is not possible to check whether a difference in age equivalent is statistically significant or to provide a confidence interval, as is possible with standardized scores (Bayley, 2006b).

These age equivalents are not included in the table for visual clarity's sake, but are very relevant in clinical practice.

3.3.2 Evaluation Form

The test administrators filled out an evaluation form immediately after they had administered the Low Motor/Vision accommodated version. In total, 12 evaluation forms out of 19 (63%) were returned by five different test administrators. The non-response was due to a lack of time on the part of the test administrators. Table 3.4 gives a summary of the responses on the three key questions of the evaluation form.

The first question pertained to whether the test results corresponded to the view of the respondent concerning the development of the child. All respondents answered positively, with six of them indicating that this correspondence was caused specifically by the Low Motor/Vision accommodations. We did not observe a difference between children with motor, visual, or motor as well as visual impairment with respect to the answer to this first question.

The second question asked whether the "Low" accommodations were practicable when testing this child. Two respondents indicated that the enlarged stimulus book was not useful, because the distance between the pictures was too large for the child to see all the pictures at once within his or her visual field. Two respondents indicated that the pictures were too dark and had too little contrast. One

Table 3.4
Responses to the Main Questions in the Evaluation Form

Question	Number of "Yes" answers	Number of "No" answers
1. Do the test results from the accommodated version correspond with your view of the developmental level of this specific child?	12	0
If "Yes", is this specifically due to the Low Motor/Vision accommodations?	6 Of which:	6
- Yes, because of accommodated materials and procedure	4	
- Yes, because of accommodated procedure	1	
- Yes, because of accommodated materials	1	
2. Were the Low Motor/Vision accommodations practicable when testing this child?		
- No, enlarged stimulus book not useful		2
- No, pictures too dark and too little contrast		2
- No, enlarged blocks were a disadvantage		1
3. What were the advantages of the Low Motor/Vision accommodations for this child when compared to the standard version? (More than a single answer possible)		
- Materials	5	
- Procedure	6	
- Successful experiences	2	
- No need to skip items	1	
- No advantage	5	

respondent indicated that the enlarged blocks caused the tower of blocks to become too high for the child to reach the top.

The third question asked what the advantages of the "Low" accommodations were for this child. Respondents could give multiple answers. Five respondents answered that the child benefitted from the accommodations to the test materials, with one of them specifically mentioning the adjusted picture book. Six respondents indicated that the accommodations to the procedures were beneficial, with two of them specifically mentioning the removal of time limits. Two respondents indicated that the accommodations had led to more successful experiences, and one respondent noted that items could now be administered that would otherwise have been skipped.

3.4 Discussion

The current pilot study focused on whether the Low Motor/Vision accommodations to the Bayley-III were practicable for the person administering the test, and suitable and beneficial for the children in the target population.

Considering the whole sample, the statistical tests revealed that the median *raw score difference* and the median *percentage score difference on adjusted items* were not significantly larger than zero. This means that the current data do not support the hypothesis that the majority of the children with a motor and/or visual impairment would obtain a higher score on the Low Motor/Vision accommodated version than on the standard version. The non-significance can be due to a lack of power – the sample size is rather small – or due to an absence of the expected score difference in the population. Even if the latter was the case, the accommodations could be beneficial for some of the children within the target population.

When considering the individuals' *raw score difference* and *percentage score difference on adjusted items*, it is salient that the variability between the children is rather large. The *raw score difference* ranges between -7 and 11 points, and the *percentage score difference on adjusted items* between -33% and 67% across the five scales. We presume that this variability is due to both differences in responses to the accommodations, and to factors as mood, health, and attention level of the child. The latter factors, typically referred to as measurement error, complicate the demonstration of structural differences between the scores on the Low Motor/Vision accommodated version and those on the standard version.

For the Cognition scale, two children (out of 19) stand out by scoring substantially higher (10 and 11 points, 25% and 43%, respectively) on the Cognition scale of the Low Motor/Vision accommodated version, compared to the standard version. This also resulted in a major difference in age equivalent scores (i.e., an increase of up to five months in developmental age). The two children both had a motor impairment and no visual impairment. One of these two children had clearly benefited from the enlarged materials, as revealed by the test report. It would be very interesting to know what caused the two children to obtain such a large score difference, but this cannot be deducted from the data in the current study.

When considering the average *percentage score difference on adjusted items* at the sample level per scale, the Low Motor/Vision accommodated version and the standard version showed about equal results for Receptive Communication (average change 1%), in contrast to the Cognition and Expressive Communication scales (average change 6% and 11%, respectively). The lack of difference in the Receptive Communication scale could be due to a negative influence from the enlarged stimulus book in combination with a positive influence from other accommodations.

The stimulus book was one of the main accommodations in the scale and appeared to be unsuitable for some children because of the large distance between the pictures and the poor contrast found in those pictures.

The responses on the evaluation form indicated that the accommodations are practical. In addition, all respondents indicated that the test results corresponded to their picture of the developmental level of the child. Half of the respondents indicated that the Low Motor/Vision accommodated version had advantages, compared to the standard version of the Bayley-III. This implies that the Low Motor/Vision accommodated version is an improvement compared to the standard version of the instrument for a subgroup of children in the target population.

We used the feedback from the respondents to adjust the Low Motor/Vision accommodated version. For example, we removed the enlarged stimulus book and the enlarged blocks, and in the manual we emphasized that the Low Motor accommodations were only beneficial for children who had motor impairments affecting the hands and/or arms. These adjustments were not applied during the current pilot study, but will be used for future study.

The results on both the test administrations and the evaluation form imply that some of the children with a motor and/or visual impairment did benefit from the Low Motor/Vision accommodations to the Bayley-III and some did not. The results on the current pilot study are consistent with those of earlier pilot studies on the Low Motor and Low Vision versions of BSID-II-NL (Ruiter et al., 2010; Ruiter et al., 2011), which found that the accommodations resulted in more valid test results and a smoother test administration. The experiences of test administrators and developmental psychologists in the current study were positive. That said, as a result of this pilot study we know that there is still room for improvement to be made to the instrument.

It is important to take into account a few issues when interpreting the results, namely the relatively small sample, the inclusion of children with a calendar age above 42 months, and the large variability in test scores within as well as between children. We minimized variability due to inconsistencies in test administrations by giving an intensive training to the test administrators. The small sample in combination with the large variability in test scores means that we are unable to draw conclusions about the effect of the accommodations on the test results in the target population. We have to keep in mind the goal of this pilot study, which was to get a first impression of the scores that assessment with the Low Motor/Vision accommodated version yields. The inclusion of children with a calendar age above 42 months in the sample should be kept in mind when interpreting the results, because not much is known about the use of the Bayley-III with these older children. However, the inclusion of the older children was important, because many children with motor and/or visual impairment who are in need of developmental assessment are older than 42 months of age and have developmental delay.

Future research should focus on the use of the Bayley-III with children older than 42 months and on the construct validity of the Bayley-III Low Motor/Vision version. The idea of increased construct validity would be supported if (a) children with a motor and / or visual impairment obtain a higher score on the Low Motor/Vision accommodated version than on the standard version, and (b) the expected value of the item score is equal for the accommodated and standard versions of the item, in so far as both versions are suitable for the child under study. The latter would imply that the norm tables of the original version still apply when the accommodations are applied.

If future research manages to develop assessment instruments that are more suitable for children with an impairment than the current set of instruments are, and research results support their validity, this will have major implications for practice. Developmental psychologists will then be able to assess the development of children with a motor and/or visual impairment more validly than is currently possible and will consequently be able to provide more adequate support.

3.5 Conclusion

The overall aim of this pilot research was to examine whether a Low Motor/Vision accommodated version of the Bayley-III is more suitable and practical than the standard version when evaluating the development of young children with a motor and/or visual impairment. In sum, the results mean that it is possible to apply the Low Motor/Vision accommodations to the Bayley-III in test administrations with children with a motor and/or visual impairment, and that the accommodations are beneficial for at least a subgroup of children within the target population.

Chapter 4

Validity and suitability of the Bayley-III Low Motor/Vision version: A comparative study among young children with and without motor and/or visual impairments*

* This chapter has been published as: Visser, L., Ruiter, S. A. J., Van der Meulen, B. F., Ruijssenaars, A. J. J. M., & Timmerman, M. E. (2013). Validity and suitability of the Bayley-III low Motor/Vision version: A comparative study among young children with and without motor and/or visual impairments. *Research in Developmental Disabilities, 34*, 3736-3745. doi:10.1016/j.ridd.2013.07.027

Abstract

The aim of the present study was to examine the validity of the Bayley-III Low Motor/Vision version, and its suitability for children with motor and/or visual impairment(s). This version contains accommodated items, that is, adaptations to minimize impairment bias, without altering what the test measures. We hypothesized that the accommodations would not affect the item scores of children without impairment, and that children with impairment(s) would benefit from the accommodations. We tested 41 children without impairment and 63 children with impairment with both the standard Bayley-III and the Low Motor/Vision versions, in randomly counterbalanced order. The test administrators filled in an evaluation form. Results showed that the accommodations did not affect the test scores of children without impairment and did improve the test scores of children with impairment on the Cognition scale, while no improvement was found for the other scales. The test administrators indicated that the vast majority of the children with impairment had been able to show their abilities on the test and that the accommodations were beneficial in 29 out of these 52 cases. For some children, the accommodated instrument appeared to be unsuitable because the impairment was too severe. The conclusion is that the accommodations improve the validity of the Bayley-III when used with children with mild to moderate motor and/or visual impairment, especially with regard to the Cognition scale.

4.1 Introduction

Development of children in the first few years of life usually happens relatively spontaneously when basic conditions in the environment, like appropriate sensory input and responsive relationships, are being met (Shonkoff, 2007). However, in cases of physical impairment, normal development cannot be taken for granted (Ziviani, Darlington, Feeney, Rodger, & Watter, 2013) and there is an increased risk for developmental problems (Hatton, Bailey, Burchinal, & Ferrell, 1997; McLaren, Edwards, Ruddick, Zabjek, & McKeever, 2011). A developmental assessment is needed in that case in order to answer questions about strong and weak areas of development (Petermann & Macha, 2008), which can then be used as a basis for early intervention to prevent secondary developmental problems (Guralnick & Conlon, 2007).

In this article we will focus on motor and visual impairments and their consequences for the developmental assessment of young children. As a consequence of the impairment(s), a child might not be able to show his or her abilities in the domain of interest during the standardized assessment of a developmental test (Hebbeler, Barton, & Mallik, 2008; Neisworth & Bagnato, 2004; Visser, Ruiter, Van der Meulen, Ruijssenaars, & Timmerman, 2012). For example, the child might not be able to manipulate the test materials, or just may need more time to deal with the material than a child without the physical impairment but with the same ability level. For the sake of objectivity, it is not advisable for test administrators to apply an unstandardized adjustment in order to overcome the influence of the impairment (Hebbeler et al., 2008). Therefore, instruments are needed that have been developed or adapted especially for children with impairment (M. R. Johnson, Wilhelm, Eisert, & Halperin-Phillips, 2001; Miller & Skillman, 2003).

The Bayley Scales of Infant and Toddler Development, Third Edition (Bayley-III; Bayley, 2006a) is a widely used and researched instrument for developmental assessment of young children. This instrument enables assessment of the levels of Cognition, Receptive Communication, Expressive Communication, Fine Motor Development, and Gross Motor Development. We have accommodated the instrument to increase the validity and suitability for use with children with motor and/or visual impairment. We have named the accommodated instrument the “Bayley-III Low Motor/Vision version” (LM/LVi version). The term “Low” refers to the number of motor and visual components in the items. Pilot research has suggested that the accommodations are advantageous for at least a subgroup of children with motor and/or visual impairment (Visser, Ruiter, Van der Meulen, Ruijssenaars, & Timmerman, in press). We could not deduce from the results of the pilot study what the characteristics of this subgroup were. Test administrators indicated that the results on the LM/LVi version provided a good picture of the development of the children and that this was specifically due to the accommodations in the majority of cases.

The idea behind the LM/LVi accommodations is to enhance the children's opportunities to show their cognitive, language, and motor skills in a test situation. We did this by accommodating the Bayley-III: we intended to lower the amount of motor and visual components in the test items, provided that the item content and difficulty would not be changed (Alant & Casey, 2005; Thurlow, Elliot, & Ysseldyke, 2003). This would imply that the construct validity of the LM/LVi version would be higher compared to the standard Bayley-III for children with motor and/or visual impairment, because the child's competencies of interest could be estimated more precisely. In the LM/LVi version, we minimized the motor component (e.g., pointing) in items designed to measure cognitive ability (e.g., connecting similar pictures) (Visser et al., in press). Other examples of accommodations made to the test are enlarged materials and increased contrast between the colors of the test materials and the background color. Obviously, the Fine Motor scale and Gross Motor scale items could not be accommodated for children with a motor impairment, because that would change the item content and difficulty.

The aim of the current study was to investigate whether the content and difficulty of the test have remained the same in spite of the LM/LVi accommodations, and whether the accommodations are beneficial for children with motor and/or visual impairment.

The first hypothesis was that the test results of children without impairment show invariant test content and difficulty. This hypothesis would be supported if two results are found. First, children have, on average, equal raw scores on the LM/LVi and the standard versions of each scale (Expectation 1a). Second, children have, on average, equal scores on the accommodated items of the LM/LVi version and their non-accommodated counterparts in the standard version, thereby correcting for possible differences in learning effects (Expectation 1b).

The second hypothesis was that the accommodations are beneficial for children with impairment. This hypothesis would be supported if three results are found. First, children have, on average, a higher raw score on the LM/LVi version than on the standard version of each scale (Expectation 2a). Second, children have, on average, higher scores on the accommodated items of the LM/LVi version than on their non-accommodated counterparts in the standard version, thereby correcting for possible differences in learning effects (Expectation 2b). Third, test administrators evaluate the LM/LVi version positively for children with a motor and/or visual impairment (Expectation 2c).

4.2 Method

4.2.1 Participants

4.2.1.1 Control group.

The control group consisted of 41 children without developmental problems. We recruited the children via convenience sampling (Gravetter & Forzano, 2012): colleagues and acquaintances of the researchers and test administrators were asked whether their children might be able to participate, and sometimes a participant's parent informed one of their acquaintances about the research as well. In addition, an invitation to participate in the research was placed on the Facebook website of one of the test administrators. The reason for choosing convenience sampling was that it was the most feasible approach, given that there is no reason to expect that the subject of interest (i.e., the difference between test scores on the LM/LVi versus the standard version) would differ among children with various individual characteristics (e.g., would depend on socio-economical, geographical, or ethnic backgrounds). There were 25 girls and 16 boys with an average age of 2 years and 15 days (range from 1 month and 18 days to 3 years, 8 months, and 20 days).

4.2.1.2 Special needs group.

The special needs group consisted of 63 children with motor and/or visual impairment who were referred by the developmental psychologist of an organization supporting children in this target group. We approached 49 branches of organizations in total, belonging to 26 different umbrella organizations that included rehabilitation centers, expertise centers for visual impairment, and organizations supporting children with learning disabilities. In total, 22 of the branches actually referred children for the current research. The branches were located in different regions of the Netherlands and, in one case, in Flanders, the Dutch-speaking part of Belgium.

We formulated four inclusion criteria. First, the child should have a diagnosis of mild to severe motor impairment affecting arm and/or hand movement, and/or a diagnosed or suspected visual impairment. Visual impairment was defined broadly, including disorders of the eye (referred to as "Low vision" by the International Statistical Classification of Diseases and Related Health Problems, Tenth Revision [ICD-10]; World Health Organization, 1992) as well as cerebral visual impairment (Dutton, McKillop, & Saidkasimova, 2006). Blind children were excluded. Second, the calendar age should be between 6 months and 11 years. Third, the presumed developmental age should be between 1 and 42 months (age range Bayley-III; Bayley, 2006a). Fourth, the child should have the minimum abilities required for the test: ability to sit upright in a chair or wheelchair, and ability to use at least one hand.

The developmental psychologist provided the information about diagnoses and impairment via a referral form. The sample included 32 girls and 31 boys, with an average age of 5 years and 12 days (range from 13 months and 4 days to 10 years, 6 months, and 18 days) at the moment of the first test administration. In Table 4.1 we summarize the information about the diagnoses and impairment of the children.

The test results were simultaneously used for our research and in the diagnostic process performed by the organization. The plan was to administer all five scales of the Bayley-III to the children with visual impairment, and only the Cognition and Receptive and Expressive Communication scales to children with motor impairment and without visual impairment. Because the Fine and Gross Motor scales contain only Low Vision accommodations, the standard and accommodated versions are equivalent for children with motor impairment and without visual impairment; comparing the standard and accommodated versions in this group would thus not be useful. In cases of both motor and visual impairments, we followed the advice of the referring psychologist in deciding whether or not it was useful to administer the motor scales.

The developmental psychologists indicated that the children who participated were representative of the target group for developmental assessment in their organization. Reasons for not referring children who did meet the inclusion criteria were: the child in question's health problems, the child having other appointments, parents not wanting to overburden their child, no response from parents, no time for referral by the psychologist involved, and administering the previous version of the instrument being preferred for comparative purposes.

4.2.1.3 Missing data and exclusion of test results.

Administered tests of seven children from the special needs group and three children from the control group were unsuitable to be included in the current study and we did not include them in the numbers mentioned in sections 4.2.1.1 and 4.2.1.2. Two children from the special needs group were only administered the accommodated version, because the second test administration was canceled and could not be rescheduled. The test results of two other children from the special needs group were entirely unreliable as indicated by the test administrator: one child did not cooperate during the test, and the other child did not have sufficient motor control in the arms and hands to be able to perform the test-item tasks. In addition, the data of three children from the special needs group and three children from the control group were not included in the study, because the data from all the subscales administered appeared to be unreliable. We defined a non-reliable scale as having more than three not-scored items per scale.

Table 4.1
Number Count of Diagnoses and Impairments of the Special Needs Group

	Impairment			
Diagnosis	Motor	Visual	Motor and Visual	*Total*
None	6	4	7	*17*
PDD	1	0	0	*1*
CP	5	0	2	*7*
Down syndrome	0	0	1	*1*
Angelman syndrome	1	0	1	*2*
Other genetic disorder	6	3	8	*17*
Other	5	0	5	*10*
CP & down syndrome	1	0	0	*1*
PDD & other	0	1	1	*2*
PDD & other genetic disorder	1	0	0	*1*
ADHD & down syndrome	0	0	1	*1*
No information	3	0	0	*3*
Total	*29*	*8*	*26*	*63*

Note. PDD = pervasive developmental disorder; ADHD = attention deficit hyperactivity disorder; CP = cerebral palsy.

All the other referred children were administered both the standard Bayley-III and the LM/LVi version. The test could not be completed for some children because of time constraints on the part of the organization or tiredness of the child. As a result, not all scales have been administered to all children. In addition, in some test administrations the results of one of the scales of the Bayley-III was not reliable as a consequence of limited cooperation on the part of the child during some part of the test session. In these cases, we excluded the results on the unreliably administered scales and only included the results on the reliably administered scales. Table 4.2 gives an overview of the number of children reliably tested with each of the scales and the average raw scores. We have not included standardized scores in this table, because many children in the special needs group were older than the limit of the age range of the Bayley-III (42 months) and in such cases a standardized score is not available.

4.2.2 Instruments

4.2.2.1 The standard version of the Bayley-III.

The Bayley-III (Bayley, 2006a) is an individually administered instrument that assesses the development of children between 1 and 42 months of age. It contains the scales Cognition, Receptive Communication (RC), Expressive Communication (EC), Fine Motor Development (FM), and Gross Motor Development (GM). Items are scored positively (1) when a child has shown the target behavior and negatively (0) when not. The starting point depends on the calendar age of the child and the starting rule is that the first three administered items must receive a score of 1. If this is not the case, the test administrator will then have to go back to the previous starting point until the starting rule has been met. Items before the starting point are not administered and are automatically scored 1. The stopping rule is to stop administering after five consecutive items have been scored 0, and all items after this final item are automatically scored 0. The domains of social-emotional development and adaptive behavior can be assessed on the basis of primary caregiver responses to a questionnaire, which was not part of the current research.

The standardization sample of the Bayley-III in the United States included 1 700 children. The internal consistency and test-retest stability appeared to be good (Bayley, 2006b). In our study we used the Dutch Bayley-III, which is identical to the American version, except for the language used (Dutch rather than English) and two deleted items in the Expressive Communication scale. Standardization research in the Netherlands is currently ongoing.

4.2.2.2 Bayley-III Low Motor/Vision version.

The LM/LVi version is similar to the standard version of the Bayley-III except for the accommodations. The scoring procedure is identical to that of the standard version. We applied the Low Motor accommodations to the Cognition and

Receptive and Expressive Communication scales, but not to the Fine Motor and Gross Motor scales, because there the impaired skill is precisely what would be measured. We applied the Low Vision accommodations to all five scales. We were able to combine the Low Motor and Low Vision accommodations into one test version. This will facilitate the application of the Bayley-III in the assessment of children with both motor and visual impairments, which is very relevant, because a large subgroup of children in the target group have motor as well as visual disabilities.

The accommodations can be categorized into material accommodations, instruction accommodations, and procedural accommodations. Examples of material accommodations are enlarged materials, addition of small handles to puzzle pieces (Low Motor), change of the color of the materials to yellow, and using a dark blue placemat as a background for optimal contrast (Low Vision). Examples of instruction accommodations are allowing elbow guidance by the test administrator when necessary (Low Motor), and added instructions to help make sure the materials are within the visual field of the child (Low Vision). An example of a procedural accommodation is the elimination of time limits. For a detailed description of the accommodations and the process leading to the LM/LVi version, we refer to Visser and colleagues (in press).

4.2.2.3 Evaluation form.

The test administrators of the children in the special needs group filled in an evaluation form for each child tested to indicate to what extent the LM/LVi version had been suitable for that specific child. The information involving the opinion of the test administrators about the suitability of the LM/LVi version for children in the special needs group is relevant, because the accommodations might have resulted in a smoother test administration even though this might not have been reflected in a difference in test score. The key questions asked were what the expected developmental level of the child was, whether the child had been able to show his or her abilities, and what role the accommodations had played. There was also space available for additional comments. The information about the expected developmental level of the child was given to the test administrator in advance by the developmental psychologist via the referral form.

4.2.3 Design and Procedure

We evaluated the accommodations to the Bayley-III using a counterbalanced within-subjects design. The counterbalancing was done by randomly assigning one of the orders at the moment of referral. In three cases, the order was determined by the physical presence of test materials belonging to the LM/LVi version. We aimed at an interval of one to two weeks between the two test sessions. This could not be achieved in all cases for practical reasons. In the control group, 20 children were tested with the standard version first and 21 children with the LM/LVi version first; the average interval was 10 days, with a range of 5-26 days. In the special needs group, 32 children were tested with the standard version first and 31 children in the opposite order; the average interval was 9 days, with a range of 2-21 days.

The tests for the current study were administered by an advanced university student in psychology or one in special needs education, or by the developmental psychologist or diagnostic worker of the referring organization. The students as well as the experts from the field received intensive training in administering the test, including information about the background and aims of the instrument, the administration and scoring of the items, and case studies. The experts from the field had experience with the previous version of the instrument and therefore the intensive training received was sufficient to be able to properly administer the test. The students carried out five test administrations in order to practice before they started to test for our research data. They video-recorded two of these test administrations and received feedback on their interaction with the child, their way of administering the test items, and their scoring. The feedback was generally limited to a couple of feedback points and no serious errors were observed. There were regular meetings with the test administrators, during which the students shared experiences and asked questions. Items that appeared to be difficult to score were discussed during these meetings.

The children from the special needs group were referred via the referral form, and the developmental psychologist asked for permission from the parents. In those cases where the developmental psychologist or diagnostic worker administered the test, they sent the record form together with the referral form, permission form, and a questionnaire about family background to the researchers. When it was one of the 18 test administrators from the university administering the tests, he/she would contact the developmental psychologist to make the appointments, and the tests were then administered at the organization which the children attended multiple days a week. A person who knew the child well was present during the test, whenever possible. In 57 out of the 63 cases the test administrator was the same across test sessions. We aimed for the same test administrator, because the psychologists indicated this to be far preferable for the sake of the children.

The children from the control group were enrolled by their parents via a short referral form. A test administrator from the university then contacted the parents. One child was tested at the university the first time and at home the second time. One child was tested at the day care center the first time and at home the second time. The other children were tested at the same location for both test sessions: 7 at the university and 32 at home. The test administrator for the control group was the same across test sessions for 39 of the 41 children.

4.2.4 Analysis

The data were analyzed using the Statistical Package for Social Sciences (SPSS) Version 20 software. In the following we refer to the Expectations described at the end of section 4.1. The numbers of cases involved in each analysis correspond to the “N” in Table 4.2 for each scale.

For our analyses, we derived three variables per scale from the items of the LM/LVi and the standard versions, namely the raw difference score (for Expectations 1a and 2a), the accommodated difference score, and the non-accommodated difference score (for Expectations 1b and 2b). The raw difference score is the difference in Bayley-III raw scores between the second (M2) and first (M1) test sessions, thus reflecting the degree of increase (or decrease) in score between the testing occasions. The Bayley-III raw score itself is important since it is the basis for the Bayley-III test result when the instrument is applied in early intervention practice.

The accommodated difference score is the difference in proportion correct of the accommodated items between M2 and M1, considering the impairment of the child. That is, we incorporated the items with Low Vision accommodations for children with only visual impairment; the items with Low Motor accommodations for children with only motor impairment; and all the accommodated items for the children with both visual and motor impairment, or without impairment. The non-accommodated difference score is the difference between M2 and M1 in proportion correct of the non-accommodated items, again considering the impairment of the child. In our results, we have presented both test statistics and 95%-confidence intervals so as to indicate the size of the effect. For all computed scores, we took into account items that were administered as well as items that were not administered, applying the Bayley-III starting and stopping rules (e.g., items before the first administered item were given a score of 1; items after the last administered item were given a score of 0).

Table 4.2
N and Mean Raw Score per Scale of the Bayley-III and per Subsample

		SN group		Control group	
		LM/LVi	Standard	LM/LVi	Standard
Cognition	N	61		40	
(91 items)	Raw score, M (sd)	43.5 (17.9)	42.2 (18.0)	59.8 (21.6)	60.0 (21.6)
RC	N	49		31	
(49 items)	Raw score, M (sd)	15.3 (9.3)	15.0 (9.0)	28.0 (13.3)	27.1 (12.8)
EC	N	52		34	
(46 items)	Raw score, M (sd)	15.8 (11.1)	15.6 (10.3)	27.6 (14.3)	27.3 (14.3)
FM	N	19		36	
(66 items)	Raw score, M (sd)	27.7 (11.1)	27.6 (11.3)	39.0 (15.3)	39.0 (14.7)
GM	N	14		29	
(72 items)	Raw score, M (sd)	40.9 (13.3)	40.6 (12.6)	46.8 (17.9)	47.5 (18.1)

Note. SN = special needs; LM/LVi = LM/LVi version; Standard = standard version; N = number of children tested with both versions of this scale, excluding non-reliable test administrations; M = mean; sd = standard deviation.

To test Expectation 1a, we performed a t-test on the raw difference score of children in the control group in order to compare the test order groups (i.e., children tested with the standard version first compared to children tested in opposite order). To test Expectation 2a, we performed the same t-test in the special needs group. We expected that a learning effect between the first and second test sessions would affect the test scores and therefore took the test order into account in our analyses.

To test Expectation 1b, we performed an analysis of covariance (ANCOVA) with the accommodated difference score of children in the control group as dependent variable, the test order as between-subject factor, and the non-accommodated difference score as the covariate. The covariate was used to control for possible differences in learning effects between the two groups of testing order. To test Expectation 2b, we performed the same ANCOVA in the special needs group.

To test Expectation 2c, we elaborately examined the results on the evaluation forms. We mapped in how many cases the child had been able to show his or her abilities, how often this was specifically due to the LM/LVi accommodations, and whether this outcome could be related to the type of impairment and/or the developmental level of the child. For the latter, we performed a chi-square test. We also studied the respondents' additional comments on the evaluation forms.

4.3 Results and Discussion

4.3.1 Test Results

4.3.1.1 Control group.

In the upper part of Table 4.3, we present the mean raw difference score of children in the control group for each scale and the mean accommodated difference score in the two test order groups. Furthermore, we show the 95%-confidence intervals of the differences in means between the test order groups. Thus, for example, children from the control group who were administered the LM/LVi version first had an average difference in raw scores between M1 and M2 of 1.4 on the Cognition scale. Children from the control group who were administered the standard version first had an average difference in raw scores of 1.0 on that scale. The difference between these two values is 0.4, with a 95% confidence interval (95% CI) of -1.6 to 2.4.

The t-tests with the raw difference score as dependent variable yielded no significant differences between children tested with the standard version first and children tested in the opposite order, for any of the scales (t = 0.41, p = 0.69 for Cognition; t = -1.98, p = 0.06 for RC; t = -0.90, p = 0.37 for EC; t = -0.08, p = 0.93 for FM; t = 1.43, p = 0.16 for GM), with a maximum width of the 95% CIs of 4.0. This suggests that both versions result, on average, in (reasonably close to) equal scores among control children, thereby providing support for Expectation 1a.

Performed per scale, the ANCOVAs with the accommodated difference score as dependent variable, the test order as between-subject factor, and the non-accommodated score difference as covariate revealed no significant differences between children tested with the standard version first and children tested in the opposite order, for any of the scales (F = 0.01, p = 0.92 for Cognition; F = 3.63, p = 0.07 for RC; F = 1.19, p = 0.28 for EC; F = 0.01, p = 0.93 for FM; F = 0.33, p = 0.57 for GM, with maximum width of the 95% CI of 0.09). This suggests that both versions result, on average, in (reasonably close to) equal scores on the accommodated part of the test among children in the control group, thereby providing support for Expectation 1b.

From those analyses, we conclude that, as expected, the accommodations do not appear to change the scores of children in the control group thereby implying that the accommodations have not changed the content and difficulty of the test items.

4.3.1.2 Special needs group.

In the lower part of Table 4.3, we present the mean raw difference score of children in the special needs group for each scale and the mean accommodated difference score in the two test order groups.

The t-tests with the raw difference score as dependent variable indicated a significant difference between children tested with the standard version first and children tested in the opposite order for the Cognition scale (t = -2.59, p = 0.01), with a 95% CI [-4.6, -0.6]. More specifically, children who were administered the standard version first showed a larger (positive) raw difference score than those children who were administered the LM/LVi version first. Because the raw difference score was calculated by deducting the score on M1 from the score on M2 (e.g., M2 minus M1), this means that the children obtained a higher score on the LM/LVi version than on the standard version, on average. This would mean that the children had benefited from the accommodations made to the Cognition scale. The t-tests with the raw difference score for the other four scales revealed no significant differences between the two test order groups (t = -0.73, p = 0.47 for RC; t = -0.80, p = 0.43 for EC; t = -0.02, p = 0.99 for FM; t = -0.93, p = 0.37 for GM). This implies a lack of evidence for a positive influence of the LM/LVi accommodations on both the raw scores on the Receptive and Expressive Communication scales for children with a motor and/or visual impairment and the Fine and Gross Motor scales for children with a visual impairment. Thus, we found support for Expectation 2a for the Cognition scale, but not for the Communication and Motor scales.

The results of the ANCOVAs with the accommodated difference score as dependent variable, the test order as between-subject factor, and the non-accommodated score difference as covariate revealed results comparable to those of the t-tests on the raw difference scores. That is, we found a significant difference between the two test order groups for the Cognition scale (F = 7.07, p = 0.01), but not for the Receptive and Expressive Communication scales and the Fine and Gross Motor scales (F = 3.59, p = 0.06 for RC; F = 1.36, p = 0.25 for EC; F = 0.65, p = 0.43 for FM; F = 0.88, p = 0.37 for GM). This suggests that the children benefited from the accommodations to the Cognition scale, but not from the accommodations

Table 4.3

Difference in Scores on the Bayley-III Standard and LM/LVi Versions: M2 minus M1

			Testing order	Cognition	RC	EC	FM	GM
Control group	Raw diff. score	Mean (sd)	1. LM/LVi 2. Stand.	1.4 (3.1)	-0.2 (2.8)	0.6 (3.3)	1.1 (2.1)	0.4 (2.3)
			1. Stand. 2. LM/LVi	1.0 (3.1)	1.5 (2.0)	1.4 (1.5)	1.1 (1.8)	-0.9 (2.5)
		95% CI of test order difference		[-1.6, 2.4]	[-3.4, 0.1]	[-2.7, 1.0]	[-1.4, 1.3]	[-0.6, 3.1]
	Acc. diff. score	Mean (sd)	1. LM/LVi 2. Stand.	0.016 (0.037)	-0.005 (0.064)	0.000 (0.110)	0.017 (0.035)	0.010 (0.040)
			1. Stand. 2. LM/LVi	0.013 (0.036)	0.033 (0.044)	0.035 (0.051)	0.018 (0.029)	-0.007 (0.042)
		95% CI of test order difference		[-0.023, 0.025]	[-0.078, 0.003]	[-0.095, 0.029]	[-0.023, 0.021]	[-0.021, 0.037]
Special needs group	Raw diff. score	Mean (sd)	1. LM/LVi 2. Stand.	-0.9 (4.0)	-0.6 (2.3)	0.2 (2.6)	0.8 (3.2)	-0.6 (1.9)
			1. Stand. 2. LM/LVi	1.7 (3.9)	-0.2 (1.3)	0.7 (1.9)	0.8 (1.9)	0.1 (0.7)
		95% CI of test order difference		[-4.6, -0.6]*	[-1.5, 0.7]	[-1.8, 0.8]	[-2.5, 2.5]	[-2.4, 1.0]
	Acc. diff. score	Mean (sd)	1. LM/LVi 2. Stand.	-0.011 (0.050)	-0.018 (0.048)	0.005 (0.058)	0.011 (0.048)	-0.023 (0.052)
			1. Stand. 2. LM/LVi	0.022 (0.048)	-0.002 (0.025)	0.024 (0.050)	0.013 (0.032)	0.000 (0.019)
		95% CI of test order difference		[-0.057, -0.008]*	[-0.044, 0.001]	[-0.048, .013]	[-0.047, 0.021]	[-0.066, 0.027]

Note. M1 = first testing session; M2 = second testing session; Raw diff. score = raw difference score; Acc. diff. score = accommodated difference score; sd = standard deviation; Stand. = standard version; CI = confidence interval.

* Significant difference between groups at $\alpha = 0.05$.

to the Receptive and Expressive Communication scales and Fine and Gross Motor scales, at least in such a way that it is reflected in the test scores. Expectation 2b can thus also be accepted for the Cognition scale, but not for the Receptive and Expressive Communication scales and the Fine and Gross Motor scales.

The sample sizes of the special needs group for the Fine and Gross Motor scales were relatively small because the LM/LVi version was not administered to children with a motor impairment. A lack of power could have played a role in the non-significant results (Peers, 1996). In addition, the number of accommodated items may have played a role in the absence of a significant score difference for the EC and GM scales. In these scales, between 55% and 60% of the items did not contain any accommodation, while between 7% and 13% of the items were not accommodated in the other three scales. For RC, the nature of the accommodations may have played a role, with many accommodations in procedure and instruction, but only a few in the materials.

4.3.1.3 Discussion of test results.

It should be noted that the variability in scores from consecutive test administrations in the current study appears to be large. This can be seen in the relatively large standard deviations of the difference scores (see Table 4.3). This large variation could be due to differences in learning effects, but also due to factors such as the tiredness, mood, and health of the child. We presume that the within-subject variation due to the latter factors is larger in our study than in daily practice. Depending on the setting, an expert might be able to try to delay testing until the child is optimally focused on the test. Due to time constraints, we were unable to wait for the most optimal moment of testing for each child in the current study. We did not include the results of those tests in which a child cooperated very poorly (unreliable result), and so we thereby filtered out the results of test administrations that contained a very large effect due to alertness variation. A recommendation for future research as well as for early intervention practice would be to regard optimal alertness of the child during testing as a necessary condition for a valid test administration.

4.3.2 Evaluation Form

The evaluation form was filled in and returned for 58 out of the 63 children in the special needs group. Table 4.4 summarizes the findings from the evaluation forms. A large majority of respondents (52 out of 58) reported that the child had been able to show his or her abilities, of which more than half (29) stated that this could specifically be attributed to the LM/LVi accommodations. The enlarged materials and the elimination of time limits appeared to play the greatest role in enabling the children to show their abilities. All other types of accommodations were beneficial for at least a few of the children. Respondents who checked “Other” indicated, for example, that the enlarged pictures and use of the eye-gaze direction of the child were beneficial. It should be noted that respondents were able to check multiple options, and so the numbers in Table 4.4 belonging to the second question do not add up to 29.

Table 4.4
Questions and Responses Evaluation Form

	What is the expected developmental level of the child?					
	Severe delay	Delay	Low average	Average	N/I	*Total*
1. Has the child been able to show his/her abilities?						
o Yes, not specifically caused by the accommodations	15	5	0	2	1	*23*
o Yes, specifically caused by the accommodations	21	6	1	1	0	*29*
o No	5	0	0	0	0	*5*
o N/I	0	1	0	0	0	*1*
Total	*41*	*12*	*1*	*3*	*1*	58
2. If yes, caused by accommodations, which accommodations?						
o Larger materials	11	5	0	1	n/a	*17*
o Adjusted colors	6	0	0	0		*6*
o Adjusted test procedure	5	0	1	0		*6*
o Adjusted item instruction	2	0	0	0		*2*
o No time limits	16	6	1	0		*23*
o Other	3	0	0	0		*3*
3. If no, what was the cause?						
o Behavior/health/concentration problems of the child	1	n/a	n/a	n/a	n/a	*1*
o Impairment of the child	1					*1*
o Characteristics of the test	2					*2*
o Environmental aspects	1					*1*
o Other	1					*1*

Note. N/I = No information; n/a = not applicable.

For the five children who were unable to show their abilities, different causes were indicated. For one child, behavior/health/concentration problems, impairment, as well as the characteristics of the test played a role, and the respondent noted that the instrument was not suitable for a child as handicapped as this child was. For the second child, test characteristics were the cause, and the respondent noted that the items in the Cognition scale still contained too many motor components. For the third child, the respondent did not indicate what the cause was. For the other two children, the impairment was not the main cause for the child not showing his/her abilities: one child had just moved to the Netherlands and interacting was problematic, while the other did not feel comfortable in the testing environment. All children who had been unable to show their abilities were characterized by severely delayed development.

We conducted two Chi-square tests to see whether the expected developmental level or the impairment of the child was related to the answers on the evaluation form. To make sure sufficient cases were present to perform the statistical test, we only took into account two categories for developmental level ("delay" and "severe delay") and impairment ("motor" and "motor and visual"). In addition, we only took into account the categories "Yes, not specifically caused by the accommodations" and "Yes, specifically caused by the accommodations" with respect to the first question stated in Table 4.4. We found no differences in the frequencies of the two answers given to this question for children with different developmental levels ($\chi^2 = 0.049$; $p = 0.824$) or impairment ($\chi^2 = 0.354$; $p = 0.552$).

Most respondents made additional comments on the evaluation form. Three of the respondents explained that the severe motor impairment of the child still prevented him/her from showing his/her abilities on the Cognition scale. Three other respondents wrote that the child had a very low developmental level and therefore the test was unsuitable. The starting point for the child's calendar age and the starting rule appeared to create problems when testing children with severe delay. Two respondents reported having to go back multiple starting points as a consequence of the severe delay in combination with the irregular scores, where items scored positively and negatively alternated more than average. Having to go back each time that the child scored negatively on a single item caused problems in these cases. In future, when using the instrument with children with severe delay, it would therefore be advisable to start with the item corresponding to the expected developmental level of the child. Another reason for the instrument being unsuitable for children with a very low developmental level is that the items on the low end of the scale measure baby behavior that is not always present in older children with severe delay. In addition, many items on the low end of the scale were not provided with accommodations and therefore the children with a low developmental level could not benefit from the accommodations as much as children with a higher level of development.

Most of the other comments gave feedback on specific materials or other accommodations. Two respondents noted that it was beneficial that they were allowed to use the eye-gaze direction instead of the child having to point to the correct picture or object. Three respondents wrote that the pen board or other enlarged materials were not beneficial because they were too large for the child in question. Four respondents indicated that the placemat that was used to work on was a distraction because it had been folded and did not lie flat. An alternative in the form of a tablecloth in the same color appeared to work better.

The findings from the evaluation form mean that the Bayley-III LM/LVi version was suitable for a vast majority of the children in the current study, and that about half of the children benefited from the accommodations. These results correspond to the results from earlier studies with the Low Motor and Low Vision versions of the Bayley Scales of Infant Development, Second Edition (Ruiter, Nakken, Van der Meulen, & Lunenborg, 2010; Ruiter, Nakken, Janssen, Van der Meulen, & Looijestijn, 2011) and the pilot study on the LM/LVi version of the Bayley-III (Visser et al., in press), which showed that most children benefit from the accommodations and that test administrators are in general positive about the accommodated version because it usually results in a smoother test administration.

The developmental level or specific impairment of the child did not explain why the child under study did or did not benefit from the accommodations. Based on the additional comments by the respondents, we have deduced that the suitability of the instrument depends on factors such as the height of the child and the size of the hands, whether spasm is present, and the preferences of the child for specific materials. We have also deduced that most of the problems with the instrument exist with children with a very low developmental level or severe motor impairment. A child needs to have the minimum of motor abilities in order to be able to conduct the activities that are part of the test items so that the test can be administered validly. Future research is necessary to develop an appropriate standardized instrument for developmental assessment for children with severe impairment and/or developmental delay, who do not have the minimum abilities required for the Bayley-III, as described in section 4.2.1.1.

A professional working with the test will have to judge whether the test is going to be suitable and which of the accommodations will be beneficial for the specific child that is going to be tested. Not all of the accommodations should necessarily be used with all of the children; only those accommodations should be applied that are expected to be beneficial for the child in question. In other words, the items of the standard and LM/LVi version are exchangeable.

4.4 Conclusion

Taken together, the results from the current study show that the LM/LVi accommodations to the Bayley-III have not changed the content and difficulty of the test items. This is reflected in the equal scores of children in the control group on the standard and accommodated versions. The results also show that children with a motor and/or visual impairment benefit from the accommodations to the Cognition scale, as shown by an increased raw score as well as an increased score on accommodated items. This benefit was not reflected in the test scores of the children in the current study when it came to the Receptive and Expressive Communication scales and the Fine and Gross Motor scales. Test administrators reported, however, that a large majority of the children in the special needs group had been able to show their abilities and, in more than half of these cases, this was specifically due to the accommodations. Although not reflected in the quantitative results of the evaluation forms, an analysis of the respondents' comments does indicate that the Bayley-III LM/LVi version might not be suitable for children with a very low developmental level and/or severe motor impairments.

These results imply that the LM/LVi version yields a more valid assessment among children with motor and/or visual impairment than the standard Bayley-III does. This conclusion is especially plausible with regard to the Cognition scale, and in cases of mild to moderate developmental delay and impairment. An instrument which enables a more valid assessment of the development of children with motor and/or visual impairment meets the demands of developmental psychologists working in daily practice with children in this target group. It means that the developmental level can be estimated more precisely. As a result, intervention can be better adjusted to the needs of the child, and hence problems secondary to the impairment can be prevented as much as possible.

Chapter 5

Low verbal assessment with the Bayley-III*

* This chapter has been submitted as: Visser, L., Ruiter S.A.J., Van der Meulen, B.F., Ruijssenaars, A.J.J.M., Timmerman, M.E. (2013). *Low verbal assessment with the Bayley-III*.

Abstract

Purpose: The purpose of the current study was to evaluate the validity of the Bayley-III-NL Low Verbal for developmental assessment of children with a language impairment. This Low Verbal version has recently been developed and consists of an accommodated Cognition scale, and non-accommodated Communication and Motor scales. *Method:* The Bayley-III Low Verbal was administered to 69 children. We compared the test results to those of a large sample of children without impairment using nonparametric Item Response Theory (IRT). We used an evaluation form and interviews to incorporate the opinion of test administrators and developmental psychologists. *Results:* The IRT analyses suggest that the test items have the same characteristics in the two different groups. The results on the evaluation form and interviews revealed that the Bayley-III-NL Low Verbal is suitable for children with a language impairment, and that the Low Verbal accommodations were beneficial for a majority of the children in the sample. *Conclusions:* We conclude that the Bayley-III-NL Low Verbal can validly assess the development of young children with a language impairment. The Low Verbal accommodations are specifically advantageous for children up to 36 months old, and appear to be suitable in cases of general developmental delay too.

5.1 Introduction

Many children receiving early intervention services have problems with their language development. Mashburn (2010), for example, reports that 62% of children in early intervention have a language impairment. By a language impairment, we mean a speech or language impairment, defined by the Individuals with Disabilities Education Improvement Act (IDEA; 2004b) as a communication disorder that adversely affects a child's educational performance. Language development appears to be interrelated with development in other areas such as cognition (Thal, 1991), especially at a very young age. This might explain why a language impairment at an early age is associated with worse outcomes in the longer term in domains such as communication, cognition, and educational attainment (C. J. Johnson, Beitchman, & Brownlie, 2010). This is especially so when a child has both receptive and expressive communication problems (McCabe, 2005); there is a high comorbidity with socioemotional and behavioral disorders.

When early language impairment comes to light, it is essential to apply developmental assessment and subsequently early intervention in order to reduce the potential negative consequences for later development as much as possible. As part of the developmental assessment, a standardized instrument is often used to map the developmental level of a child in different domains. The administration of a standardized measurement should always be combined with other methods (Individuals with Disabilities Education Improvement Act, 2004a) such as observation, background information, and interviews with other professionals and parents. In addition, the choice for a specific standardized instrument should be made for each child individually in order to assure that the instrument is suitable. As Crais (2009) points out, a standardized instrument is only fair when its validity and reliability are supported, and the test has representative norms. For example, the results of tests assessing development in domains other than language should not be confounded by the child's language impairment. However, available standardized instruments for developmental assessment appear to be not always suitable in cases of language impairment.

Therefore, nonverbal instruments are often used to assess cognitive development when a child is diagnosed with or expected to have a language impairment. For children with a calendar age between 2;6 (2 years and 6 months) and 7;0 years, the *Snijders-Oomen Nonverbal Intelligence Test Revised* (SON-R-2½-7; Tellegen, Winkel, Wijnberg-Williams, & Laros, 1998b) is available in Czech, Dutch, English, French, German, and Slovakian (Tellegen & Laros, 2011). For children with a calendar age between 0;1 and 2;6 years, the nonverbal version of the Dutch *Bayley Scales of Infant Development* (2nd ed.) (BSID-II-NL; Ruiter, Hoekstra, Van der Meulen, Lutje Spelberg, & Nakken, 2006) is available. In actual practice, the BSID-II-NL is also being used for children older than 2;6 years who have an expected developmental age of less than 2;6 years. Thus, the SON-R-2½-7 and nonverbal BSID-II-NL complement each other for the various calendar and (presumed) developmental age groups. For a review of instruments for developmental assessment of young children that include a nonverbal component, we refer to Visser, Ruiter, Van der Meulen, Ruijssenaars, and Timmerman (2012).

The current study focuses on the *Bayley Scales of Infant and Toddler Development* (3rd ed.) (Bayley-III; Bayley, 2006a). This is the most recent successor of the BSID-II (Bayley, 1993; Van der Meulen, Ruiter, Spelberg, & Smrkovsky, 2002). The Bayley-III has improved test materials and an adjusted structure for the developmental scales. The psychometric qualities appear to have been improved, and updated norms are available (Bayley, 2006b). It is a widely used and researched instrument for developmental assessment of children between 0;1 and 3;6 years of calendar age. As was the case with the BSID-II, the Bayley-III is also being used with older children who have a suspected developmental age of less than 3;6 years. The instrument assesses development in the areas of Cognition, Receptive Communication, Expressive Communication, Fine Motor Development, and Gross Motor Development.

We developed a Low Verbal version of the Bayley-III-NL, the Dutch version of the Bayley-III. We did this by accommodating the Bayley-III-NL; our objective was to lower the degree of verbal ability needed to perform the test items, while keeping the item content and difficulty unchanged (Thurlow, Elliot, & Ysseldyke, 2003). This would then imply a higher construct validity for children with a language impairment. Furthermore, the standard norms would remain applicable.

Because the Cognition scale of the Bayley-III-NL does contain language elements, while the Cognition scale aims to measure cognition and not language, we accommodated the test items. These accommodations involved the general procedure as well as the item instructions, and were based on the BSID-II Nonverbal accommodations. These accommodations are intended to lower the number of language components in the test items and thereby empower the child with a language impairment to show his or her cognitive abilities. The term “Low” refers to the number of language components in the test items.

We did not accommodate the Communication scales of the Bayley-III-NL, because these assess the language skills of a child. Lowering the number of language components in the items would therefore change the content of the scale and thus would threaten the construct validity. We also did not accommodate the Motor scales, because the items belonging to the Motor scales rely to a very limited extent on language, and therefore we expected that accommodations would not be necessary.

This paper describes the results of a validity study on the Bayley-III-NL Low Verbal. The main question was whether the accommodated Cognition scale and the non-accommodated Communication and Motor scales provide a valid assessment of the development of young children with a language impairment. To answer this question, we compared the item responses of children in the target group on each scale of the Bayley-III-NL Low Verbal with those of children without developmental problems on the standard version of the instrument. To this end, we used item response analyses. The expectation was that the items would have the same characteristics in the two different groups and thus could be sensibly used in both groups. Moreover, we used an evaluation form and interviews with developmental psychologists to assess to what extent the Bayley-III-NL Low Verbal is more suitable than the standard version for assessing the development of young children with a language impairment.

5.2 Method

5.2.1 Participants

The study contained two samples: a special needs group of children with a language impairment and a control group.

5.2.1.1 Special needs group.

The inclusion criteria for the special needs group were:

- The child has a language impairment.
- Calendar age between 0;6 and 11;0 years.
- Expected developmental age between 0;1 and 3;6 years, which is the developmental age range of the Bayley-III.
- The minimum abilities required for the test: ability to sit upright in a chair or wheelchair; ability to use at least one hand; exclusion of children with severe

visual impairment so test materials could be well perceived; and exclusion of children with deafness.

This sample consisted of 69 children: 26 girls and 43 boys. The mean age was 4;8 years with a range of 2;2 to 10;8 years. The children were referred by 17 different branches of organizations, and all referred children were tested. The organizations included those supporting young children with speech and/or language problems specifically, as well as those supporting a more broadly defined group of young children with special needs. We received information about nonresponse from six of the 17 organizations via evaluative phone calls. Three of these six organizations reported that there were between one and four children who could have been referred, but were not. The reasons reported were difficulty communicating with parents as a consequence of a language barrier, parents' resistance to testing, health problems of the child, and the child moving to a different organization.

The developmental psychologist from the referring organization provided information about the (suspected) presence of general developmental delay and the (suspected) diagnosis and impairment of the child. Among the 69 children, 59 had (suspected) general developmental delay, eight had an expected average cognitive development, and for two children we did not have information about the developmental level in advance. We have summarized the information concerning diagnoses and impairment of the children in Table 5.1. The language impairment was always the primary impairment of the child, which was also true for the children who also had a motor and/or visual impairment.

5.2.1.2 Control group.

The inclusion criteria for the control group were that the child should have an (expected) development within the normal range, no known impairments, and a calendar age between 0;1 and 3;6 years. The control group consisted of 1132 children from different parts of the Netherlands. The average age was 1;7 years, with a minimum of 0;1 and a maximum of 3;6 years. There were 547 girls and 585 boys. The control group consisted of all children who were included in the sample of the Bayley-III-NL standardization research up to April 11, 2013. The standardization study will run from 2010 to early 2014.

Table 5.1
Diagnoses and Impairments: Special Needs Group, Main Study

	Impairment				
Diagnosis[a]	LI	LI & Motor	LI & Visual & Motor	Unknown	*Total*
None	29	3	0	0	*32*
PDD	9	4	0	0	*13*
Down syndrome	5	2	1	0	*8*
ADHD	2	0	0	0	*2*
Angelman syndrome	0	1	0	0	*1*
Other genetic disorder	2	0	1	0	*3*
Other	0	4	1	0	*5*
Hydrocephalus & other	0	1	0	0	*1*
PDD & other	1	0	0	0	*1*
Other genetic disorder & other	0	1	0	0	*1*
Unknown	1	0	0	1	2
Total	*49*	*16*	*3*	*1*	*69*

Note. LI = language impairment; PDD = pervasive developmental disorder; ADHD = attention deficit hyperactivity disorder.
[a] The numbers include children who had already received a diagnosis as well as children who were still in the assessment process.

5.2.2 Instruments

The children from the control group were tested with the standard Bayley-III-NL; the children from the special needs group were tested with the Bayley-III-NL Low Verbal, which consists of the accommodated Cognition scale and the non-accommodated Communication and Motor scales.

5.2.2.1 Bayley-III-NL.

The Bayley-III (Bayley, 2006a) is an individually administered instrument for developmental assessment of children in the calendar age range of 0;1 to 3;6 years. It is also being used with older children who have an expected developmental age within this range. The instrument assesses the development in the areas of Cognition (91 items), Receptive Communication (49 items), Expressive Communication (48 items), Fine Motor development (66 items), and Gross Motor development (72 items). The scales are built up of items increasing in difficulty. Items are scored dichotomously: a child receives a score of 1 when he/she passes an item and 0 when he/she fails an item. The starting point depends on the calendar age of the child. Items before the starting point are not administered and automatically scored 1. Regarding the basal and ceiling rules, the basal rule is that the first three administered items should have a score of 1. If not, the test administrator should go back to the previous starting point until the rule has been met. The ceiling rule is to stop after five consecutive items have been scored 0; all items after the final administered item are not administered and automatically scored 0. A raw score can be calculated for each scale, counting all items that scored 1. For children with a calendar age of up to 3;6 years, a standard score (average 10, sd 3) and index score (average 100, sd 15) can be calculated on the basis of this raw score. For children with a calendar age of more than 3;6 years, a developmental age equivalent can be looked up for each scale, which corresponds to the raw score.

The domains of social-emotional development and adaptive behavior can be assessed on the basis of primary caregiver responses to a questionnaire, which was not part of this current research. The standardization sample of the Bayley-III in the United States included 1700 children. Validity data are given in the form of moderate to high correlations of Bayley-III test scores with scores on other instruments. The internal consistency and test-retest stability of the Bayley-III in the United States appeared to be good (Bayley, 2006b).

The Bayley-III-NL is different from the American version in terms of language used (Dutch rather than English) and two deleted items in the Expressive Communication scale, resulting in a total of 46 items. The two items were deleted, because they appeared to be unsuitable with respect to Dutch language development.

5.2.2.2 Bayley-III-NL Low Verbal: characteristics and development.

The Bayley-III-NL Low Verbal has the same characteristics as the standard version of the instrument, except for the accommodations. The manual and the scoring form differ from the standard Bayley-III-NL, whereas the test materials are the same. The accommodations to the Cognition scale are based on those of the BSID-II-NL Nonverbal version (Ruiter, Hoekstra, Van der Meulen, Lutje Spelberg, & Nakken, 2006) and can be divided into a procedural accommodation and item accommodations. The procedural accommodation means that the test administrator should administer the items with emphasis: using short sentences, talking clearly, and placing extra emphasis on the key words when talking to the child. This is described in the Low Verbal manual. The item accommodations include the addition of commonly used signs to items in which language comprehension of the child is necessary in the standard version. An example is the thumbs-up gesture to say “Well done!” An instruction was also added to applicable items describing how to attract the attention of the child. These signs and instructions are described explicitly in the Low Verbal manual per item, as an addition to the item instructions of the standard version of the instrument, providing an optimal degree of standardization. The scoring form contains a pictogram before each item that contains a Low Verbal accommodation so that the test administrator is reminded to take the accommodations into account. When a commonly used sign is applicable, the scoring form provides a pictogram or short word referring to that particular sign so that the test administrator does not need the manual during testing.

We named the first version of the accommodated instrument that we developed the “Nonverbal” version. We performed a pilot study in which we assessed the suitability of the Nonverbal version for children in the target group. The sample consisted of 11 children with a language or hearing impairment. The remaining inclusion criteria were the same as described above. The children in the pilot study were tested with both the Bayley-III-NL Nonverbal and the standard Bayley-III-NL, in randomly assigned counterbalanced order, so that the test administrator could compare the two testing sessions. The test administrators filled in an evaluation form about the suitability and practicality of the accommodations after each test administration with the Nonverbal version. In addition, semi-structured interviews were performed with two experts who worked as developmental psychologists with children with a language or hearing impairment.

The results of the evaluation form showed that the accommodations were suitable and practical in the majority of cases (10 out of 11 cases). For three children, the suitability was reported not to be specifically caused by the accommodations. These children had limited communication abilities due to pervasive developmental disorder and, as a consequence, did not have sufficient communication abilities to acknowledge the gestures. The seven children for whom the accommodations were reported to be beneficial obtained an equal or higher score on the accommodated version as compared to the standard version (0 to 3 raw score points higher), in contrast to the child for whom the Bayley-III-NL Nonverbal was not suitable, who scored 1 raw score point lower. This child was deaf, and the accommodations were just not sufficient to compensate for this.

The outcome of the two interviews confirmed this last point. The two experts indicated that the Bayley-III-NL Nonverbal is not suitable for children above 3 years of age who are deaf, because the test items higher on in the scale still contained too much language. They also indicated that the Nonverbal Cognition scale still contained too much language to be called a nonverbal test. The experts did find the accommodated version suitable for children with a language problem, however.

During the pilot study, it appeared that administering the standard Bayley-III-NL to a child with a language or hearing impairment was not useful. The test administrators regularly indicated that a very awkward situation arose when they tried to administer the standard Bayley-III-NL to a child from the target group, because it came across as very unnatural and insensible when they restrained from using supportive gestures while they were actually needed. They also reported occasionally and unintentionally incorporating (unstandardized) adjustments in the standard test administration, such as the use of gestures.

On the basis of the results from the pilot study, we adjusted the accommodated instrument and planned the main research. We added information to the scoring form and manual. For example, we added pictures of the gestures and a note that the items in the higher age range (above 3 years) still require a verbal answer from the child. We renamed the instrument “Low Verbal.” For the main study, we decided to administer only the Low Verbal version and not the standard version to the children. Furthermore, we decided to concentrate the main study on the suitability of the Bayley-III-NL Low Verbal for children with a language impairment, because the instrument appeared to be especially suitable for children in this group.

5.2.2.3 Instruments for qualitative evaluation of the Bayley-III-NL Low Verbal.

The test administrators in the main study filled in an evaluation form for each child in the special needs group. They were asked about the suitability of the Bayley-III-NL Low Verbal for the specific child and about possible errors in the material.

To obtain an overview of the opinion of developmental psychologists about the suitability and added value of the Bayley-III-NL Low Verbal, a structured interview was done with four developmental psychologists who worked with young children with a language impairment. These interviews were conducted after the tests for the current research were completed. All the psychologists had already had experience with the Bayley-III-NL Low Verbal by observing one or more test administrations with the instrument on a child they knew, and by interpreting the results of that test administration. One respondent worked at an organization supporting young children with developmental disabilities and their families, while three of the respondents worked at an organization supporting young children with a language impairment. The respondents had been working as developmental psychologists for between six months and ten years.

5.2.3 Procedure

5.2.3.1 Special needs group.

We received approval for the research from the Medical Ethical Committee of the University Medical Center Groningen, located in the Netherlands.

The referring developmental psychologist filled in a referral form with information about the child, such as expected developmental level, diagnoses, and impairments. Permission was asked from the parents via a consent form. The child was then tested at the organization, which was a familiar location, because the child attended it multiple days a week. Two children were tested by the referring psychologist; the other children were tested by a test administrator from the university. These test administrators were advanced university students in special needs education or psychology. Both the students and the psychologists who administered the tests for this study had received training in administering the standard and Low Verbal versions of the Bayley-III-NL. The training program included one 4-hour session, during which the participants received information about the instrument and completed a case example of a test administration. The students also carried out five test administrations with children from their own circle of acquaintances in order to practice. They recorded two of these five test administrations and received feedback from one of the researchers, who was very experienced with the Bayley-III-NL.

After the test administration, the test administrator filled in the evaluation form and wrote up a test report. The report and test results were given to the referring psychologist so that they could be used in the regular developmental evaluation of the child.

Four structured interviews were performed with developmental psychologists who had been involved in the research via one or more of the test administrations. Two of those interviews were carried out by one of the researchers, and two by two different test administrators. The interview questions were answered in written form by the respondent before the actual interview, and the interviewer then complemented these written answers during and immediately after the interview with information obtained by asking additional questions.

5.2.3.2 Control group.

The tests of the children in the control group were administered as part of the standardization research being done in the Netherlands, which is being carried out by our colleagues from Utrecht University. Separate approval was obtained for this research via the Medical Ethical Committee of the University Medical Center Utrecht. Parents of children who met the inclusion criteria were approached using flyers and information letters via multiple channels, including infant welfare centers, child day-care centers, and local governments across the entire Netherlands. Parents of children who agreed to participate in the research by returning a consent form were invited for a test administration. Some children were tested at home (n=176), but most children were tested at their day-care center or a different location outside their home, such as a testing room in an arranged location near the child's home.

5.2.3.3 Exclusion of non-reliable results.

The aim was to administer all five scales of the Bayley-III-NL to all children in the special needs group as well as in the control group. Because the Bayley-III is an extensive test with an average 90-minute duration when given to older children, it appeared not always possible to administer all the scales. In addition, we categorized all the scales with more than 3 not-scored items (e.g., refused by the child or forgotten by the test administrator) as non-reliable. This resulted in the following numbers of reliable test administrations per scale. For the special needs group, 67 children were reliably administered the Cognition scale, 58 the Receptive Communication, 57 the Expressive Communication, 60 the Fine Motor development, and 48 the Gross Motor development scale. No children were excluded, because none of the children had a non-reliable administration for all the administered scales. For the control group, 1073 children were reliably administered the Cognition scale, 1063 the Receptive Communication, 1088 the Expressive Communication, 1078 the Fine Motor, and 1066 the Gross Motor development scale. For four children, all the scales appeared to have been administered non-reliably. We excluded these children when calculating the total number of children in the sample (n = 1132), and the corresponding average and range of the calendar age.

5.2.4 Analysis

The analyses for the main study consisted of a quantitative part, associated with the test results, and a qualitative part, associated with the evaluation form and interview.

5.2.4.1 Quantitative.

We started with a descriptive analysis of the test scores of the children in the special needs group and in the control group. We summarized the mean, standard deviation, and the range of the test scores per scale, using the scoring rules of the Bayley-III. Furthermore, we considered the number of times each item was actually administered. It was important to take this into account when interpreting the results, because an item analysis partly based on imputed values – albeit according to the scoring rules of the Bayley-III – might yield biased results. For each item, in each target group, we required an absolute minimum of 20 observed scores in order to draw any conclusion on that specific item in the target group involved.

To analyze the item scores, we used nonparametric Item Response Theory (IRT). IRT can be used to investigate the properties of a set of items, including underlying traits and the reliability of the resulting scale. Basically, it uses the Item Response Function (IRF), which is the success probability for an item depending on the latent trait, in order to study the characteristics of items and a set of items. Specific response probabilities for each specific person-item combination are modeled (Sijtsma & Molenaar, 2002). In nonparametric IRT models, data do not have to fit a specific parametric form of the IRF. This means that having to force data into a structure that they sometimes do not have can be avoided, resulting in information that is relatively easy to interpret (Meijer & Baneke, 2004). Mokken scale analysis is a well-known nonparametric IRT analysis, based on the Monotone Homogeneity Model (MHM; Mokken, 1971).

The Bayley-III is a comprehensive and complex instrument. This complexity increases the chance of data not fitting a parametric model. Therefore, we studied the item characteristics of each of the five scales of the Bayley-III-NL using Mokken scale analysis in MSP5 (Molenaar & Sijtsma, 2000). We made a data file with the dichotomous item scores, applying the scoring rules of the Bayley-III to comply as much as possible with the way the instrument is used in practice. The scores of the children per item were thus imputed, where all items before the first administered item were given a score 1 (basal rule), all items after the last administered item were given a score 0 (ceiling rule), and not-scored items (items that were not scored although they should have been scored) were given a score 0.

We ran a Mokken scale analysis in MSP5 separately for each of the Bayley-III scales Cognition, Receptive Communication, Expressive Communication, Fine Motor development, and Gross Motor development, comparing the special needs group and the control group. We used the results to assess the fit of the MHM. We tested the assumption of monotonicity (i.e., that the IRFs are monotonically nondecreasing), and the strength of the items and scale by examining the scalability coefficients of the items (H_i) and the total scale (H). Following the criterion as set by Mokken (1971), an H_i-value and H-value should be larger than .3 to ensure that the item and the scale concerned, respectively, meet the assumptions underlying the MHM, and have sufficient discriminating power. Note that the imputation of scores according to the basal and ceiling scoring rules of the Bayley-III inflates the scalability coefficients. Therefore, we consider this H-value as the absolute minimum.

To examine the assumption of monotonicity, we examined the criterion values, given by MSP as the result of the check of monotonicity. A criterion value exceeding 80 is considered a strong indication that an item violates the assumption concerned (Molenaar & Sijtsma, 2000). In cases of violations, we inspected the empirical IRF (which is estimated on the basis of the observed data) to examine the source of the non-monotonicity.

To assess whether the item characteristics were different for the special needs group than for the control group, we looked at differential item functioning (DIF) between the two groups. DIF refers to the situation in which differences exist in the way a test item functions across certain groups that are matched on the attribute measured by that item (Osterlind & Everson, 2009). In the current study, DIF between the special needs group and the control group could originate from structural differences in item functioning between the groups; in the case of the Cognition scale, DIF could also stem from differences between the items of the Low Verbal and standard versions, due to the accommodations. If DIF is found, this would indicate that the item has different characteristics when assessed with children with a language impairment and might therefore not be suitable for that target group. We examined the criterion values given by MSP as a result of the check of equal item-step ordering across groups, using the value of 80 as a threshold for deciding whether DIF exists with regard to an item.

5.2.4.2 Qualitative.

We examined the results on the evaluation forms in great detail. We mapped the number of times the test administrators had reported that the child had been able to show his or her abilities, and how often they reported that this was specifically due to the Low Verbal accommodations. This opinion of the test administrators about the suitability of the Low Verbal version for the children in the special needs group is relevant, because it gives information about the added value of the accommodations. This information cannot be extracted from the test results in the current study, which only provided information about the comparability of the test scores to scores of children without an impairment.

To incorporate the view of the developmental psychologists working in the field, we examined the results of the structured interviews. We categorized the different kinds of answers given, mapped the frequencies of the answers, and summarized the most important results.

5.3 Results

5.3.1 Test Results

We have summarized the descriptive statistics of the test scores of the special needs group in Table 5.2, and those of the control group in Table 5.3. The tables show the number of items that each scale contains, and the number of children who were reliably assessed with that scale. The tables also show the average, standard deviation, and the range of the raw scores of the children for each scale. There appears to be a larger variation in test scores within the control group, compared to the special needs group, as reflected in the standard deviations. This difference can be explained by the range in calendar age for the special needs group of 2;2 to 10;8 years: Unlike the control sample, the special needs sample does not include very young children, resulting in relatively few low scores.

Table 5.2 also shows the number of times each item was administered to one of the children in the special needs group. The following items met the criterion of a minimum of 20 administrations: items 60-86 (Cognition scale), items 25-41 (Receptive Communication scale), items 20-40 (Expressive Communication scale), items 30-59 (Fine Motor development scale), and items 48-71 (Gross Motor development scale). In the control group, all items in all scales met the minimum of 20 administrations.

Table 5.2
Administered Items and Raw Test Scores: Special Needs Group

Number of times each item was administered [a]			Cognition (91 items)	RC (49 items)	EC (46 items)	FM (66 items)	GM (72 items)
	0-9 times	Item nos.	1-33, 37-39, 44, 89-91	1-9, 48-49	1-9	1-25, 63-66	1-43
		N [b]	40	11	9	29	43
	10-19 times	Item nos.	34-36, 40-43, 45-59, 87-88	10-14, 16-18, 20-24, 42-47	10-19, 41-46	26-29, 60-62	44-47, 72
		N [b]	24	19	16	7	5
	20-29 times	Item nos.	61-62, 82-86	15, 19, 25-27, 38-41	20-26, 28, 36-40	30, 32-34, 53-59	48-53, 67-71
		N [b]	7	9	13	11	11
	30-39 times	Item nos.	60, 74-81	30-37	27, 29-35	31, 35-37, 42, 49-52	54-56, 60-66
		N [b]	9	8	8	9	10
	≥ 40 times	Item nos.	63-73	28-29	-	38-41, 43-48	57-59
		N [b]	11	2	0	10	3
N (sample)			67	58	57	60	48
Test score		M (sd)	65.6 (13.6)	27.2 (10.0)	27.1 (10.6)	43.2 (9.4)	57.1 (9.5)
		Range	26 – 86	9 – 45	7 – 45	22 – 62	23 – 69

Note. RC = Receptive Communication; EC = Expressive Communication; FM = Fine Motor development; GM = Gross Motor development; Item nos. = Item numbers; N (sample) = number of children tested with this scale, excluding non-reliable test administrations; M = mean; sd = standard deviation.
[a] Number of times the item mentioned has actually been administered to a child in the Special Needs sample.
[b] N = Number of items mentioned in the "Item nos." row.

Table 5.4 summarizes the results of the Mokken scale analysis for the special needs group and the control group, separately. The results given are the H_i-value of the items per scale and the H-value of each scale of the Bayley-III, which were all larger than .3. In addition, the items that were reported to contain a violation of the assumption of monotonicity are reported in Table 5.4. The numbers between brackets are the criterion values that were given by MSP in cases of such a violation.

Table 5.3
Raw Test Scores: Control Group

		Cognition (91 items)	RC (49 items)	EC (46 items)	FM (66 items)	GM (72 items)
N (sample)		1073	1063	1088	1078	1066
Test score	M (sd)	50.3 (22.9)	21.1 (12.0)	22.4 (14.0)	33.1 (14.7)	42.6 (18.8)
	Range	2-86	2-46	0-46	1-65	3-71

Note. RC = Receptive Communication; EC = Expressive Communication; FM = Fine Motor development; GM = Gross Motor development; N (sample) = number of children tested with this scale, excluding non-reliable test administrations; M = mean; sd = standard deviation.

Table 5.4
Results of Mokken Scale Analysis

		Cognition	RC	EC	FM	GM
Special needs group	H_i of items: range	0.79-0.99	0.50-0.96	0.77-0.99	0.63-1.00	0.52-0.97
	H of scale	0.94	0.86	0.91	0.83	0.79
	Item with vi. mon. (crit.)	-	19 (109)	-	-	66 (69)
Control group	H_i of items: range	0.51-1.00	0.69-0.97	0.88-0.99	0.77-0.99	0.82-0.99
	H of scale	0.84	0.94	0.97	0.94	0.97
	Item(s) with vi. mon. (crit.)	1, 2, 5, 46 (83 – 141)	49 (87)	-	1 (129)	1 (195)

Note. RC = Receptive Communication; EC = Expressive Communication; FM = Fine Motor development; GM = Gross Motor development; vi. mon. = violations of the assumption of monotonicity; crit. = (range of) criterion value(s).

For Cognition, we found a violation of the assumption of monotonicity for four out of the 91 items (numbers 1, 2, 5, and 46) with respect to the control group. The criterion values for these items ranged between 83 and 141. Inspection of the empirical IRFs belonging to these four items revealed that the violations were due to values that were based on a sample of 16 children in the case of items 1 and 2, and 25-27 children in the case of the other two items.

For the Receptive Communication scale administered to the children in the special needs group, a violation of the assumption of monotonicity was indicated for one item (item 19, criterion value = 109). This item was administered to 22 children. Inspection of the empirical IRF revealed that children with a rest score between 9 and 17 had a slightly higher average score on this item than children with a rest score between 18 and 25. Inspection of the data matrix revealed that the scores of most children were in line with their rest scores, and that the violation was caused by deviant scores on this item of approximately four children. For the control group, the assumption of monotonicity appeared to be violated for item 49, the last item in the scale. Inspection of the IRF and the data matrix revealed that this was caused by the fact that only one child out of the 1063 received a score of 1 on this item, and that this child did not have the highest raw score possible, namely 41.

For the Expressive Communication scale no violations of the assumption of monotonicity were indicated for any of the two groups.

For Fine Motor development, the results of the control group showed a violation of the assumption of monotonicity for item 1 (criterion value = 129). Inspection of the empirical IRF and data matrix revealed that this was caused by the fact that only seven children obtained a score of 0 on this item, and that the raw scores of these children ranged from 5 to 12, while 29 children in the control group obtained a raw score of 4 or less.

For Gross Motor development, the results of the special needs group on item 66 appeared to violate the assumption of monotonicity. Inspection of the IRF and the data matrix revealed that this was probably caused by the results of one child, who obtained a raw score of 55, but did obtain a score of 1 on item 66. The other children with a positive score on this item all had a raw score of 63 or higher. For the control group, the results for item 1 violated the assumption of monotonicity. This appeared to be caused by the score of one child, who obtained a score of 0 on item 1, but a raw score of 9 on the Gross Motor scale (while 60 of the 1066 children in the control group obtained a lower raw score).

No Differential Item Functioning between the special needs group and the control group was indicated for any of the scales of the Bayley-III.

5.3.2 Evaluation Form

Table 5.5 summarizes the results on the main question in the evaluation form. For 40 out of the 69 children in the special needs group, the evaluation form was returned by the test administrator, and in 37 out of these 40 cases the main question was answered. The test appeared to be suitable in 33 cases, and in 10 of these cases the test administrator reported that this was specifically due to the Low Verbal accommodations. Respondents could indicate whether the accommodations in item instructions or test procedure had played a role, or whether another factor had, and the respondent could check multiple answers. Nine respondents reported that the accommodated item instructions were advantageous, and two respondents reported that the accommodated test procedure (e.g., extra emphasis on key words) was advantageous.

Four respondents reported that the child had not been able to show his/her abilities during the test. Multiple answers could be checked as to the cause. Three respondents reported that it was caused by the impairment of the child, and two respondents reported that it was caused by behavior, health, or concentration problems of the child. Three of the four children had a severely delayed development. Two of the four children had a pervasive developmental disorder or signs thereof, and one of the other two children had a motor impairment (together with the language impairment).

The respondents could leave additional remarks on the evaluation form. Based on these remarks, we filtered out two main reasons for the Low Verbal accommodations not being advantageous. The first was that the child only had an expressive language problem (mentioned five times). The second was that the child scarcely responded to his or her environment, and the gestures did not make any difference in this regard.

Table 5.5
Number of Response Options Chosen to Main Question in Evaluation Form: "Has the Child Been Able to Show his/her Abilities on the Test?"

Response option	N	
Yes, caused by Low Verbal accommodations	10	
- Adjusted item instructions		8
- Adjusted test procedure		2
- Other, namely…		0
Yes, not caused by Low Verbal accommodations	23	
No	4	
- Caused by behavior/health/concentration problems child		1
- Caused by impairment child		2
- Caused by behavior/health/concentration problems & impairment child		1
- Caused by characteristics instrument		0
- Caused by environmental factors		0
No answer	3	
Total	*40*	

5.3.3 Interviews

The interview consisted of six questions, which were the basis for the structured interviews. In the following paragraphs, we have summarized the answers to the questions and added, between brackets, the number of respondents who reported the stated issue.

The first question was: "To what extent do you think the Bayley-III-NL is suitable for the target group of your organization?" The question resulted in remarks about the Bayley-III-NL in general, which included the opinion that the instrument is suitable, especially for children between 2;0 and 2;6 years of age, and children with challenging behavior for whom the SON-R-2½-7 is not yet suitable (reported 2x). Disadvantages were the fact that Bayley-III-NL results are often not accepted as a basis for school referrals (2x), and the fact that the instrument does not yield information about scholastic skills (1x).

The second question was: “What is your opinion about the practicality and the added value of the Low Verbal accommodations?” The practicality of the gestures as well as the extra emphasis on key words was assessed by the respondents as moderate (1x) or good (3x), and for the shortened sentences as good (4x). The added value was assessed as moderate (3x) or good (1x) for the gestures; and moderate (2x) or good (2x) for the shorter sentences, and for the extra emphasis on key words. The respondents explained that the gestures are clear but give less information to the child than the linguistic instruction does, and therefore do not give sufficient support (2x), especially from item 70 onwards (1x). The Bayley-III-NL Low Verbal still contains rather long sentences (1x). The shorter sentences and extra emphasis are not sufficiently helpful in cases of low verbal abilities (1x). One respondent remarked that the gestures are logical and natural, and provide good support.

The third question was: “Does the Bayley-III-NL Low Verbal have added value, compared to the standard version, for the target group of your organization?” The respondents answered “a little” (1x) or “yes” (2x) with respect to the test result, and “yes, clearly” (3x) with respect to the course of the test administration. One respondent did not answer this question.

The fourth question was: “For which children does the Bayley-III-NL Low Verbal specifically have added value, compared to the standard version?” The respondents reported that the Low Verbal is especially suitable for children who do not have too large a delay in their receptive language development, and for whom the SON-R-2½-7 is not yet suitable. For the other children, they would prefer the SON-R-2½-7 (2x). The Low Verbal is especially suitable in cases of limited concentration abilities (1x), suspected developmental delay (1x), and speech-language problems with little active language (1x).

The fifth question was: “How did the course of the administration of the standard Communication and Motor scales of the Bayley-III-NL go, and to what extent did the results give a good picture of the development of the children in these areas?” The standard assessment of the Communication and Motor scales appeared to give a fairly reliable picture of the development of the child (3x). The results on these scales were insightful (1x), and the scales are child-friendly (1x).

The sixth question was: “What do you find are the advantages and disadvantages of the Bayley-III-NL Low Verbal when compared to the SON-R-2½-7?” The advantages are that the test materials are more appealing for the young children, compared to the book-based materials of the SON-R-2½-7 (5x), and that the results of the Bayley-III-NL give more information that is relevant in the subsequent support of the child, and in tracking the child’s development (1x). The longer duration was mentioned as an advantage, because the moments of resistance are informative (1x). The Bayley-III-NL has a more suitable abstraction level and is therefore more reliable, especially for children between 2;6 and 3;0 years of age (1x), and the results are less dependent on earlier experiences with the test materials (1x). It is more the type of test that the Bayley-III is, than the Low Verbal accommodations specifically, that lead to an added value compared to the SON-R-2½-7.

The disadvantages mentioned were that it is difficult to make a classification of the speech-language impairment for older children in the age range, based on results of the Bayley-III-NL (1x). In addition, it is sometimes difficult to explain that it is Cognition that is measured with the Bayley-III-NL (1x). The Bayley-III-NL does not measure scholastic skills and does not have a clear link with the scholastic skills that are taught during early intervention (1x). The Low Verbal accommodations do not give sufficient support for children with a severe receptive language impairment (2x). One respondent reported that the Bayley-III-NL Low Verbal is preferable for the calendar age range of 2;6 to 3;0 years, but not for older children.

5.4 Discussion

The current study focused on the suitability of the Bayley-III-NL Low Verbal for the assessment of young children’s development in cases of language impairment. We compared the test results of children with a language impairment on the accommodated Cognition scale, and on the standard Communication scales and Motor scales, to those of children without an impairment on the standard Bayley-III-NL. We asked the test administrators and four developmental psychologists whether the Bayley-III-NL Low Verbal is suitable for children with a language impairment.

The results of the Mokken scale analysis showed that the H_i-values of all items and the H-values of the scales, for both subgroups, were higher than .3. This suggests that the items meet the monotonicity assumption and that the scales sufficiently distinguish between children with different levels of the latent trait.

The check of monotonicity revealed a few items, for which the assumption of monotonicity was violated. With respect to the Cognition scale, the assumption was violated for four items in the control group, two of which were the first two items. The violations were minor, however, because they were based on the test results of relatively few children (n = 16-27, depending on the item). The fact that no violations were found in the special needs group for the Cognition scale could be due to the relatively small sample in the special needs group. The first 21 items in this scale were not administered to any child in this group but were given a score of 1 automatically on the basis of the basal rule of the Bayley-III. Consequently, any violations of monotonicity were impossible to detect in this range of items for the special needs group.

With respect to the Receptive Communication scale, the results on one item in each subgroup violated the assumption of monotonicity. Both were caused by the results of a very small number of children (n = 1-4). With respect to Expressive Communication, no violations were found. With respect to Fine Motor development and Gross Motor development, we found a violation of the assumption for the first item in the scale with respect to the control group, and one other item with respect to the special needs group. Again, inspection of the IRFs and data matrices revealed that the violations were caused by deviating test results of a small number of children (n = 1-7).

No DIF was found between the special needs group and the control group for any of the items in any of the scales. This suggests that the items do not have substantially different characteristics when assessed with a child with a language impairment than when assessed with a child without an impairment. This conclusion can only be drawn for the items that were actually administered to a considerable number of children in the special needs group: items 60-86 (Cognition scale), items 25-41 (Receptive Communication scale), items 20-40 (Expressive Communication scale), items 30-59 (Fine Motor development scale), and items 48-71 (Gross Motor development scale).

The results on the evaluation form reveal that the test administrators found the Bayley-III-NL Low Verbal suitable for a vast majority of the children, but that the Low Verbal accommodations were only beneficial for some of the children. The results of the structured interviews with developmental psychologists confirmed that the Low Verbal accommodations do not have added value for all young children with a language impairment. More specifically, the developmental psychologists reported that the accommodated instrument is especially suitable for children who do not have too large a delay in their receptive language development and for whom the SON-R-2½-7 is not yet suitable. They preferred the Bayley-III-NL Low Verbal for children up until about 3 years of age, and reported that it is also suitable for children with general developmental delay. The respondents reported that the Bayley-III-NL has clear advantages, when compared to the SON-R-2½-7, in terms of the suitability of the test materials. In addition, they reported that the unaccommodated Communication and Motor scales gave a reliable picture of the development of the children.

Taken as a whole, the results mean that the Bayley-III-NL Low Verbal Cognition scale, and the standard Communication and Motor scales, can validly assess the development of young children with a language impairment. The Low Verbal accommodations are specifically advantageous for children up to 3 years of age and for children with more general developmental delay. For older children with a severe receptive language impairment, the SON-R-2½-7 is more suitable. On a child-by-child basis, it is up to the developmental psychologist to judge which instrument is the most suitable. This recommendation is in accordance with the results of a study into the "Low Motor/Vision version" of the Bayley-III-NL, accommodated for children with a motor and/or visual impairment (Visser, Ruiter, Van der Meulen, Ruijssenaars, & Timmerman, 2013c). The Low Motor/Vision version also appeared to be suitable for a subgroup of children in the target group, and the advice with respect to that version was similar: the developmental psychologist must judge which version is the most suitable for a specific child.

A limitation of the current study is that the added value of the accommodations could not be evaluated on the basis of the test results of the children. This was caused by the fact that it appeared to be impossible to administer the standard Bayley-III-NL to children with a language impairment, because that led to a very awkward testing situation. As a consequence, we depended on the opinion of the test administrators and developmental psychologists to evaluate the suitability. From a clinical perspective, this opinion is valuable. Another limitation is the relatively small sample of the special needs group. The effort it took to recruit the children in this sample showed that it would not be feasible to obtain a larger sample. However, a sample of 69 children is in fact not large enough as a basis for Item Response analysis. In addition, the scoring rules of the Bayley-III resulted in a situation where some of the items were administered to only a few of the children or even to none of the children at all. The scores were, in these cases, based on the basal and ceiling rules of the Bayley-III. As a consequence, the H-values found in this study could, for example, be an overestimation, while the conclusions on the basis of the results of the MSP analyses can only be made for a range of the items per scale. However, it was especially the easier items, those at the low end of the scales, which were not assessed in a sufficient number of children. This was probably caused by the fact that the easier items were administered to very young children, and at that stage a language impairment might not yet be identified. The items about which we can draw no conclusions thus might not be used very often in clinical practice with children with a language impairment. Another limitation is that the indicated absence of DIF between the special needs and control group could be due to the limited sample size. This might be more likely for minor forms of DIF that would end up only barely affecting the test scores, and thus would not be relevant from a clinical perspective; however, the presence of DIF that remained unidentified in this study cannot be completely ruled out.

A strength of the current study is the fact that the test results of the special needs group could be compared to a very large control group, which was part of the standardization sample in the Netherlands. Another strength was that developmental psychologists who work with young children with a language impairment have been involved in the development as well as the evaluation of the accommodations.

Future research should aim at continuously improving the assessment possibilities for young children with a language impairment. In addition, an appropriate developmental assessment instrument is needed for young children who are deaf. One of the experts interviewed for the pilot study of the Bayley-III-NL Low Verbal proposed translating the Bayley-III-NL into a sign language version. Another option would be to make an entirely nonverbal version of the Bayley-III-NL by removing all items that contain language. The target group, however, would be relatively small, because nowadays most children with deafness receive a cochlear implant at a young age. The target group of young children with a language impairment is much larger, and we would like to once again stress the importance of the availability of a standardized and valid instrument for developmental assessment of children in this target group. Because early language impairment is associated with problems in other developmental areas in the short as well as the long term (C. J. Johnson et al., 2010; McCabe, 2005), it is essential that the development of the children in the target group can be assessed and tracked in different domains. The intention of the current study was to contribute towards reaching this goal.

Chapter 6

Dynamic assessment with the Bayley-III among young children with developmental disabilities*

* This chapter has been submitted as: Visser, L., Ruiter S.A.J., Van der Meulen, B.F., Ruijssenaars, A.J.J.M., Timmerman, M.E. (2013). *Dynamic assessment with the Bayley-III among young children with developmental disabilities.*

Abstract

This article describes the results of a pilot study into the newly developed dynamic version of the Bayley-III Cognition scale. We studied the value added to the standard Bayley-III using test administrations among 57 children with developmental disabilities and expert interviews with 6 educational psychologists. Results showed that there is clear variability in the responses of children to the help-steps that belong to the dynamic procedure, and in the score difference between the pretest and posttest. The educational psychologists indicated that the dynamic procedure clearly has added value for specific groups of children, such as children in the preliminary stage of school placement and children from socially disadvantaged families. The Task behavior questionnaire was evaluated positively by most educational psychologists. We conclude that the dynamic procedure provides added value for the developmental assessment of children with developmental disabilities. We finish with suggestions for future research into the dynamic Bayley-III.

6.1 Introduction

In the last three decades there has been growing interest in dynamic assessment of the cognitive abilities of young children with developmental disabilities. Unlike static testing, dynamic testing implies intervening with the child during the assessment so that the child is able to improve test performance, and so that the administrator can identify the amount and type of assistance that a child needs to develop his/her full potential (Haywood & Lidz, 2007; Resing, 2006). Dynamic Assessment (DA) has been applied with different clinical and educational groups, and was found to be more accurate in reflecting children's learning potential than static tests, especially with minority and learning-disabled children (R. J. Kahn, 2000; Tzuriel, 2000). DA makes it possible to analyze how modifications in testing conditions affect individual differences in test performances (Carlson & Wiedl, 1992).

DA was originally based on Vygotsky's theory of the zone of proximal development (Vygotsky, 1978) and Feuerstein's theory of mediation in learning (Feuerstein, Klein, & Tannenbaum, 1991). DA has been advanced by several authors (Grigorenko & Sternberg, 1998; Lidz, 1987; Tzuriel, 2001). The general idea is that assessment should focus on what a child can achieve with help, rather than on unassisted performance. DA is described as an interactive procedure that systematically and objectively measures the degree of change that occurs in response to cues, feedback, or adapted task conditions (Haywood & Lidz, 2007).

DA provides the antidote to two important limitations of static assessment. First, static assessment focuses on quantifying the degree of disability, providing a baseline for intervention evaluation. It does not assess learning capacities and sensitivity to instruction, and consequently does not provide information that can be used as a basis for intervention planning (Bagnato, Neisworth, & Pretti-Frontczak, 2010; R. J. Kahn, 2000; Snow & Van Hemel, 2008; Visser, Ruiter, Van der Meulen, Ruijssenaars, & Timmerman, 2012). DA does provide this information (Hessels-Schlatter, 2002; Lauchlan & Elliot, 2001; Tzuriel, 2000). Consequently, it is evaluated more positively by practitioners than static assessment, when it comes to understanding learning difficulties and planning intervention (Freeman & Miller, 2001). Second, static tests may underestimate learning capacities, because the underlying assumption is that children have had equal opportunities to acquire the skills that the tests measure. This assumption is not necessarily met in cases of a disadvantaged background, immigration, or developmental problems (Sternberg &

Grigorenko, 2001). Consequently, adherence to static tests may limit test validity. An inferior performance could be interpreted as reflecting delayed cognitive functioning, while in fact the child's performance has been reduced by external factors that have hindered full employment of his/her learning capacities.

Many DA instruments have been developed for school-aged children. A DA instrument for preschool children needs to have different characteristics, because meta-level thinking and verbal reasoning are not sufficiently developed to be able to recount problem-solving strategies. A young child expresses his/her understanding and performance level via actual behavior. Therefore, behavioral observations form the basis for the DA test result in young children, showing the degree to which the child has benefited from help, and the type and amount of assistance the child needed.

A number of DA instruments are available for children between three and six years of age (Tzuriel, 2000), and for infants and toddlers (R. J. Kahn, 2000). However, dynamic procedures are infrequently used in early intervention practice, at least in the Netherlands. Therefore, we have developed a dynamic test procedure that is an easy complement to a standardized, norm-referenced instrument that is widely used in clinical practice: the Bayley Scales of Infant and Toddler Development, Third Edition (Bayley-III; Bayley, 2006a). The aim was to adapt the testing conditions to children with developmental disabilities, who generally need more time and help in order to understand the testing situation and tasks (Bagnato, Macey, Salaway, & Lehman, 2007), thus enhancing the possibilities for using the test results in intervention planning. The idea behind the dynamic Bayley-III is that the well-supported quantitative scores of the standard version can still be used, while adding qualitative information in the form of learning responses to the help-steps, and differences in responses between the pretest and posttest.

The current article describes the results of a pilot study into the use of the dynamic Bayley-III with young children with developmental disabilities. The research question was whether the dynamic procedure has added clinical value when administering the Bayley-III to children with developmental disabilities. We hypothesized that (a) the dynamic test procedure yields information about sensitivity to instruction that differs among children, (b) the questionnaire on Task behavior differentiates between children, and (c) educational psychologists will evaluate the Bayley-III more positively when the dynamic procedure is added.

6.2 Materials and Methods

6.2.1 Participants and Procedure

This study is embedded in a broad research project focusing on effective and fair developmental assessment of young children with cognitive and/or functional disabilities, and was approved by the Medical Ethical Committee of the University Medical Center Groningen, the Netherlands. We used test administrations and expert interviews for evaluating the dynamic Bayley-III.

6.2.1.1 Children.

The sample consisted of 57 children with developmental disabilities (25 girls, 32 boys; age: *M* = 3;11; *SD* = 2;4; range 1;00 - 9;10). All children were known to a center for assessment and intervention for persons with developmental disabilities in the Netherlands. The test results could be integrated into the regular treatment structures of the centers. The educational psychologists referred children who (a) were diagnosed with developmental disabilities; (b) had a developmental age range between 1;00 and 3;6 years, with a maximum calendar age of 10;00 years; and (c) had sufficient auditory, visual, and fine motor abilities to accomplish the test tasks. The sample included children with pervasive developmental disorder, cerebral palsy, hydrocephalus, Down syndrome and other genetic disorders, motor, visual and/or speech-language impairment, and children with no diagnosis other than developmental disabilities.

The children were tested at the referring organization. The mean time between the pretest and posttest was 7 days (range 4 to 14 days). Ten children were tested by the referring educational psychologist; the other children were tested by one of 14 university students in special needs education or psychology. All administrators had received intensive training and were supervised by the researchers. Six children had a different test administrator for the pretest and posttest.

The Task behavior questionnaire was returned for 93 children with developmental disabilities (age: *M* = 3;9 years; range 0;8 to 9;11) from the sample of the broad research project, 38 of whom were from the sample of the current study.

6.2.1.2 Interview participants.

After finishing all test administrations for the current study, the educational psychologists were invited to evaluate the adequacy and added clinical value of the dynamic Bayley-III via an interview. Six educational psychologists from five different organizations volunteered. Two psychologists from the same organization preferred to have one interview together, leading to a total of five interviews. All participants were educational psychologists working with young children with developmental disabilities. One participant had personally administered the dynamic Bayley-III about 15 times, one participant had observed multiple test administrations with the dynamic version, and four participants had only read the dynamic test reports.

For each participant, we developed a dynamic test report based on a test administration with a child who had been referred by that participant. The case report and interview questions were sent out in advance to facilitate preparation for the interview. The interviews were administered by one of the researchers and a Master’s student who was under close supervision of the researchers.

6.2.2 Instruments

6.2.2.1 Bayley-III standard.

The Bayley-III (Bayley, 2006a) is an individually administered instrument that assesses the development of children between one and 42 months of age. The instrument consists of the scales Cognition, Receptive communication, Expressive communication, Fine motor development, and Gross motor development. In this study we used the Bayley-III-NL, which is identical to the American version, except for the language used (Dutch rather than English) and two deleted items in the Expressive communication scale. Standardization research in the Netherlands is currently ongoing (completion in 2014). The items are dichotomously scored. The set of items administered depends on the child’s age (entry point) and the developmental level of the child (applying reverse and discontinue rules). Raw scores can be calculated for each subscale by adding up the number of passes and the number of unadministered items below the basal level. The raw score of the Cognition scale was used in the current study.

The standardization sample in the United States included 1700 children. Bayley-III test scores correlated moderately to highly with scores on other relevant instruments. The reliability and test-retest stability appeared to be good (Bayley, 2006b).

6.2.2.2 Bayley III dynamic.

The dynamic procedure was designed for the Cognitive scale and consists of a pretest-training-posttest format. It involves administering the standard version of the Cognition scale (pretest), immediately followed by a training session on the negatively scored items. The pretest and training session together take about one hour. The training session helps the child to gain a better understanding of the task and to build feelings of competence. A week after the pretest and training session, the Cognition scale is administered again, as a posttest. The Cognition scale consists of 91 items, and therefore the maximum value of the raw score is 91. The dynamic procedure has been designed for items 31 to 91 (i.e., age range 12-42 months), implying that the procedure is suitable for children with an estimated developmental age between 12 and 42 months.

The training for negatively scored items consists of a few or all of the four help-steps for each item, gradually increasing in terms of the amount of help provided. (1) *Repetition* involves repeating the instruction for the item in the standard version. (2) *Explanation* means the test administrator provides verbal support to the child during completion of the task. (3) *Demonstration* involves the test administrator physically showing the child how to complete the task. (4) *Joint completion* means the test administrator completes the task together with the child, for example, by using hand-over-hand guidance. The manual describes, per item, which help-steps are applicable and how these should be administered, thus aiming for an optimal level of standardization. The test administrator applies the help-steps until the child is able to accomplish the task, and notes the final help-step needed and the responses of the child. The manual describes how the test scores should be summarized and interpreted, thereby focusing on the child's responses for the trained items.

6.2.2.3 Task behavior questionnaire.

We developed a questionnaire for mapping the task-related behaviors of the child during test administration, such as concentration, flexibility, impulsivity, perseverance, and alertness. The questionnaire is a modification of several Dutch instruments on the same subject (Veenstra, Van Geert, & Van der Meulen, 2010), consisting of 12 questions referring to behaviors of the child, which are scored on a 5-point Likert scale ranging from "never" (1) to "often" (5). The questions were answered by the test administrator and by a person who knew the child well, shortly after the Bayley-III test administration.

6.2.2.4 Interviews.

The interviews were semi-structured and supported by a topic list and a dynamic test report. The latter was based on a dynamic test administration with a child who had been referred by the participant. Each report was written on the basis of the guidelines in the manual concerning how to summarize and interpret the test scores.

The topic list was developed on the basis of the research questions and consisted of the following open questions:

1. Does administering the dynamic Bayley-III have added value for (part of) the target group of your organization, when compared to administering only the standard Bayley-III?
2. What do you think about the practical usability of the dynamic procedure?
3. Do the test results of the dynamic procedure yield additional information that is useful for an intervention plan?
4. What is your opinion about the dynamic test report?
5. What is your opinion about the extra time needed for the dynamic procedure, compared to using the standard version only?
6. What is your opinion about the Task behavior questionnaire?

6.2.3 Analyses

6.2.3.1 Test results.

Test results of 11 children were excluded from the analysis, because of unreliable pretest, training, or posttest data, leaving 57 participants whose data was analyzed. We defined unreliable as having more than three not-scored items (e.g., refused by the child) during either pretest, training, or posttest.

The quantitative analyses consisted of computing summary measures of the test results, that is, of the raw pretest scores (T_1), the raw posttest scores (T_2), and the raw differences scores (T_2-T_1). We plotted the frequency of each of the final help-steps for each child.

6.2.3.2 Task behavior questionnaire.

We performed an exploratory ordinal common factor analysis on the results from the questionnaire to identify whether different domains could be distinguished and the resulting scales differentiated among the children. We used Parallel Analysis to indicate the number of factors, Unweighted Least Squares as the method of extraction, and oblique rotation using the Promin criterion. All analyses were performed with the software program FACTOR (Lorenzo-Seva & Ferrando, 2006). The criterion for linking questions to a particular factor was that the loading in absolute value was > .3. For the scales thus established, we computed summary measures.

6.2.3.3 Interviews.

The responses given during the interviews were written out, summarized, and coded. The coding consisted of giving a positive or negative value to the answers, and was done by both the Master's student and one of the researchers, to ensure objectivity. The researcher made the final decision about the coding in cases of disagreement. We rated a question of the topic list positively if a majority of the respondents had answered the question positively.

6.3 Results

6.3.1 Test Results

The raw score summary measures are as follows: at T_1: M = 60.7; SD = 12.6; range 29 – 80; at T_2: M = 63.4; SD = 11.5; range 38 – 82; difference score: M = 2.8; SD = 3.4; range -7 – 14; approximately normally distributed. A difference score between 0 and 4, indicating equal or improved performance, was achieved by 87.7% of the children. Figure 6.1 shows the frequency per final help-step needed to accomplish the item and the difference score, per child. There is a large variation in number of trained items as well as in responses to the help-steps: for some children, training in most cases did not result in a positive score, while for other children "repetition" or "Explanation" was generally sufficient.

6.3.2 Task Behavior Questionnaire

Factor analysis of the answers given by a person who knew the child well did not yield an interpretable set of factors, and we decided to refrain from further consideration of this part of the questionnaire.

Factor analysis of the answers by the test administrators yielded three distinguishable and interpretable factors, namely Perseverance (one question), Impulsivity (four questions), and Alertness (four questions). We left out three items from the scale, because one question could not be linked to a factor and two questions were unrelated in content to the other questions in the factor.

Figure 6.1 Responses to Help-steps

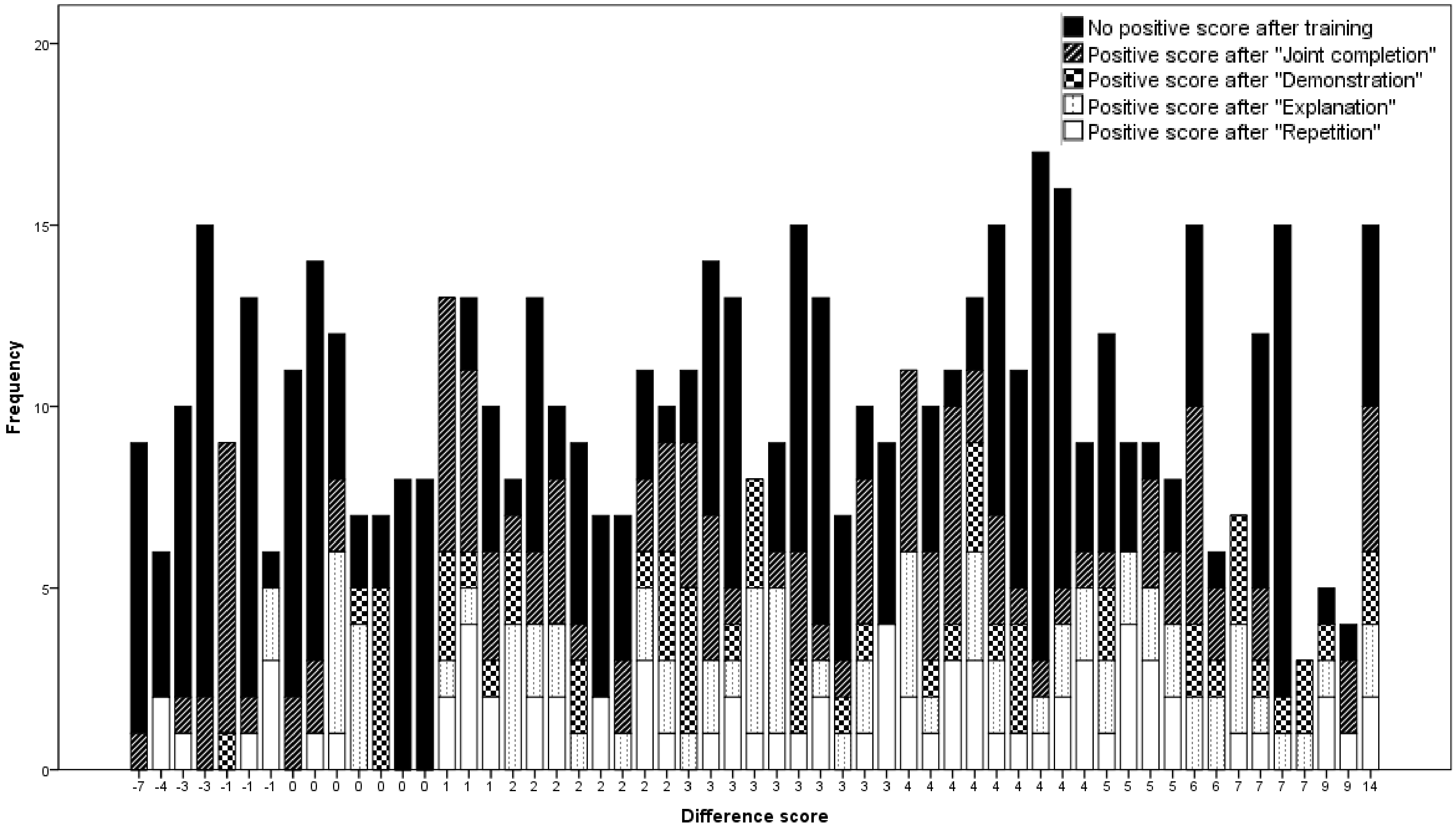

Figure 6.1. Per child, the frequency per final help-step needed to accomplish the item, and the difference score. The black parts of the bars represent the number of items, for which the child failed to achieve a positive score. The total length of each bar indicates the total number of trained items.

We defined three scales as the average of the questions linked to the factor concerned. The summary measures for the scales were as follows: Perseverance: $M = 3.7$, $SD = 1.0$; range 1 – 5; Impulsivity: $M = 2.8$, $SD = 0.9$, range 1.25 – 5; Alertness: $M = 3.1$, $SD = 0.4$; range 2 – 4. The average score on Perseverance is thus relatively high, and the scores on all three scales were approximately normally distributed.

6.3.3 Interviews

Table 6.1 summarizes the responses that were given during the interviews. We considered the interview with two respondents together as one response, making a total number of five respondents. All questions were answered positively by a majority of the respondents.

(1) *Added value of the dynamic procedure for the organization's target group.* All respondents found that the dynamic procedure clearly has added value. Respondents mentioned that the two testing moments, and the information yielded about learning potential and type of help the child benefits from have added value: A child feels more at ease in a second session and consequently shows more of his/her skills.

(2) *Practical usability of the dynamic procedure.* All respondents were predominantly positive about the practical usability of the procedure, the help-steps, and the two testing moments. Two respondents indicated that two testing moments are difficult to plan and would therefore apply the dynamic procedure only when it has a clear advantage. Respondents mentioned specific target groups for the dynamic procedure, such as children with a functional impairment, in the preliminary stage of school placement, or
from socially disadvantaged families. Opinions were divided with respect to the suitability for children with profound and multiple learning disabilities.

(3) *Information yielded by the dynamic procedure that can be used as a basis for an intervention plan.* The respondents were positive and referred to developmental level, task behavior, learning potential, help needed, and communication abilities as the different aspects of useful information obtained by the dynamic procedure. One respondent illustrated why the responses to the help-steps are informative: "A child with profound learning disabilities will mostly need the help-step of 'Joined completion,' while a child in early intervention who is developing towards a regular school placement will probably need a decreasing number of help-steps."

Table 6.1
Summary of Responses to the Interview Questions

	Positive remarks	n	Negative remarks	n
1. Added value ++*: 3* +*: 2*	- Two test moments is an advantage	4	- Questions remain	3
	- It yields valuable information about:		- Second testing moment only of moderate value	2
	• learning potential	4	- It is not right to repeat a test this soon	1
	• response to help-steps	3	- T2 is not useful if the child does not pay attention during training at T_1	1
	• way of learning, needs	2	- Daily carers already know what support the child needs	1
	- Better view of the child	2		
	- We can now legitimately give the necessary help during testing	1		
2. Practical usability +*: 5*	- Help-steps are usable	3	- 2 testing moments not feasible	2
	- Practical usability is good	2	- Intensive, costs a lot of time	1
	- Test procedure is usable	2	- Test administration takes too long	1
	- 2 testing moments feasible	2	- Scoring procedure is not practical	1
	- Manual and scoring form are user friendly	1	- Tester might automatically adjust help-steps to suit the child	1
			- Hand-over-hand causes resistance in some children	1
3. Info intervention ++*: 4* +*: 1*	- You can specify the intervention plan more	3	- In case example, many of the dynamic results were not used for the intervention plan	1
	- Meets the need for info about different aspects	3		
	- Very nice for developing support tips	2		
	- Help-steps are very informative	2		
	- Info about what child can learn is very informative	1		
4. Test report ++*: 5*	- I recognize the child in the test report	3	- The report raises questions	1
	- Processing of scores positively evaluated	3		
	- Description of item level is informative	2		
	- It provides more info than the standard version	1		
	- Test report is clear	1		
	- Elaborate description is good, important	1		

(Table continues)

Table 6.1 ***(continued)***

	Positive remarks	n	Negative remarks	n
5. Extra time ++*: 3 /* +*: 2*	- It is worth the extra time investment	3	- Worth extra time if aim of assessment indicates the use of the dynamic version	2
6. Task behavior ++*: 4* *– : 1*	- Provides guidelines for behaviors to pay attention to during observation	2	- Contains good questions, but they are part of my observations anyway; I do not need them	1
	- Concrete, practical, added value	1		
	- Yields more complete picture of child	1		
	- Results are easy to link to support tips	1		

Note. The numbers and terms in the left column refer to the six interview questions. The information printed in italics refers to the number of respondents who gave only positive (++)/mainly positive (+)/only negative remarks (–).
n = number of respondents who made the remark mentioned. Info = Information.

(4) *The case dynamic test report.* The respondents were all very positive about the way the test results had been processed and described in the report.

(5) *Extra time needed for the dynamic procedure.* Three respondents indicated that the dynamic procedure is in any case worth the extra time investment. Two respondents said that it is worth it when there is a clear reason for using the dynamic procedure.

(6) *The Task behavior questionnaire.* Four of the five respondents evaluated the questionnaire positively.

6.4 Discussion

As Lauchlan and Elliot (2001) mentioned, dynamic testing is intuitive and attractive due to its interactive nature and fine-tuning to the individual child. It is also complex, and there is an ongoing debate as to the validity and applicability of dynamic test results. This study evaluated the clinical value of the dynamic Bayley-III when used with children with developmental disabilities. We hypothesized that (a) the dynamic procedure would yield information about sensitivity to instruction that differs among children, (b) the Task behavior questionnaire would differentiate between children, and (c) educational psychologists would evaluate the Bayley-III more positively when the dynamic procedure is added.

The test results show that the children's responses varied substantially and that the help-steps in the dynamic procedure each represented a unique addition, confirming hypothesis (a). The children in the study showed substantial variation in the scores on the three scales of Alertness, Perseverance, and Impulsiveness, thus confirming hypothesis (b). The interviewed educational psychologists evaluated the dynamic Bayley-III for children with developmental disabilities positively. They particularly appreciated the practical usability and the extra information on learning potential, task behavior, communication abilities, and instructional needs that could be used in intervention planning. Critical remarks mainly concerned the extra time investment, and the practicability of the scoring and interpretation procedure. We advocate applying the dynamic procedure only when clear advantages are to be expected, such as for children in the preliminary stage of school placement, or from socially disadvantaged families, which can be identified using the 10 risk factors enumerated by Sameroff (2009). Overall, the interview results supported the claim of enhanced clinical validity due to the DA procedure, and supported hypothesis (c).

Many studies on dynamic testing focus on a score that expresses learning potential. With dynamic testing in very young children with disabilities, we deem determining a learning potential score as not useful. It is unspecific and difficult to interpret due to the large variability in everyday functioning and low predictive value of test scores in the target group (Petermann & Macha, 2008). Instead, it is important to find out what skills are emerging and what type of help the child prefers. This information can be used for designing an intervention program.

Two factors should be taken into account when interpreting the current study's results. The first is the size and diversity of the sample, which limits the generalizability of the research findings. Large variability in diagnosis and developmental level is a recurring problem in research involving children with disabilities. Even children with the same diagnosis can experience great differences in their everyday functioning, which often fluctuates due to poor health and susceptibility to environmental influences. In this study, we have been very intent on taking into account special circumstances during testing, as far as permitted by the standardized procedures. The Task behavior questionnaire is a useful tool for determining the validity of any one specific administration, because it maps behavioral factors that may have influenced the results.

The second factor is the absence of a control group that would be tested at pretest and posttest, but without the training session. This would be needed in order to answer the question to what extent the learning effect contributes to the improvement of the score at posttest. The answer should come from future, more quantitatively oriented research. Future research should also focus on the consequences of the dynamic procedure for construct and criterion validity. When the definition of cognitive development includes not only “unassisted performance” but also a dynamic component focused on “ability to learn,” it can then be claimed that DA enhances construct validity. Regarding criterion validity, it is expected that the posttest score more accurately predicts future school achievement than the pretest (static) score (Tzuriel, 2000). Finally, it would be important to investigate whether the specific target groups mentioned during the interviews do indeed benefit from the use of dynamic testing.

Chapter 7

General Discussion

The aim of the current thesis was to design and evaluate a developmental assessment instrument that is more suitable for young children with developmental impairments and disabilities than available instruments are. We designed a developmental assessment instrument in the form of a Special Needs Addition to the Bayley-III-NL. This addition consists of a dynamic procedure that yields results that are useful for intervention planning, a Low Motor/Vision version for children with visual and/or motor impairments, and a Low Verbal version for children with a speech/language impairment. The latter two were accommodated versions of the standard Bayley-III-NL.

In what follows, we will summarize the main results of the studies described in the current thesis (Section 7.1) and thereby answer the research questions that we formulated in the Introduction. This will be followed by a discussion of the limitations and strengths of the studies (Section 7.2), implications and suggestions for early intervention practice (Section 7.3), and suggestions for future research (Section 7.4).

7.1 Summary of Main Results

The project began with a literature review (Chapter 2) that focused on the question "Which instruments for developmental assessment of young children are available and what is known about their suitability in cases of special needs?" Our first focus was on issues encountered with respect to standardized developmental assessment. The results showed that four issues are regularly described in the literature: low reliability in cases of low levels of developmental functioning, dependence of test results on a child's functional limitations, test results that are suboptimal for intervention planning, and test duration that is too long. Based on the results of a systematic literature review, we concluded that a variety of standardized instruments are available for children below 4 years of developmental age and that the characteristics of these instruments vary substantially in terms of the four issues mentioned. Specific instruments for young children with motor or hearing/language impairments are available, but their psychometric properties need to be researched and improved. For young children with a visual impairment, there appeared to be no appropriate standardized instrument available at all.

The subsequent studies aimed to solve two of the issues described in the review article, namely the dependence of test results on a child's functional limitations and the suboptimality of test results for intervention planning. With respect to lowering the dependence of test results on a child's functional limitations, we formulated the research question: "How can we accommodate the Bayley-III-NL to increase the suitability for children with a visual, motor, or speech/language impairment?" We used accommodations, which means that our intent was to change the test in such a way that the influence of impairments on the test results would be lower, without changing the content of the test. In that way, the norms of the standard version could be used in the accommodated versions.

We combined the accommodations for motor and visual impairments into one version: the Low Motor/Vision version. The results of the studies, in which we evaluated this version (Chapters 3 and 4), suggest that the accommodations have not changed the content and difficulty of the test items. This is reflected in the equal scores of children in the control group on the standard and Low Motor/Vision versions. Children with a motor and/or visual impairment appeared to benefit from the accommodations to the Cognition scale, as shown by a higher score on the Low Motor/Vision version. This benefit was not reflected in the test scores with respect to the Communication and Motor scales. Test administrators reported that a large majority of the children in the special needs group had been able to show their abilities and, in more than half of these cases, this was specifically due to the accommodations. The test administrators also indicated that the Low Motor/Vision version might not be suitable for children with a very low developmental level and/or a profound motor impairment. The results imply that the Low Motor/Vision version yields a more valid assessment among children with motor and/or visual impairment than the standard Bayley-III-NL does, especially with regard to the Cognition scale, and in cases of mild to moderate developmental delay and impairment.

In the Low Verbal version, we accommodated the Bayley-III-NL for children with a speech/language impairment. The results of the study, in which we evaluated this version (Chapter 5), suggest that the test items of the Low Verbal version, administered to children with a speech/language impairment, have the same characteristics as those of the standard version when administered to children without impairment. This means that the Low Verbal version can validly assess the development of young children with language impairment. With respect to the added value of the accommodations, test administrators and developmental psychologists who had been involved in the test administrations for the study indicated that the accommodations are specifically advantageous for children younger than 3 years of age and children with general developmental delay. When developmental psychologists were asked in an evaluative interview to compare the Bayley-III-NL Low Verbal to a nonverbal intelligence test for young children, the SON-R-2½-7 (Tellegen, Winkel, Wijnberg-Williams, & Laros, 1998b), they evaluated the SON-R-2½-7 as more suitable for assessing cognitive development, especially in cases of severe receptive language impairment and a developmental age of older than 3 years.

With respect to the suboptimality of test results for intervention planning, we formulated the research question "How can we adjust the Bayley-III-NL so that the test results become more useful as a basis for intervention planning?" We developed a dynamic version that serves as an addition to the Cognition scale of the Bayley-III-NL. The results of the pilot study, in which we evaluated the clinical utility of the dynamic version (Chapter 6), showed that the children's responses to the help-steps and Task behavior questionnaire, both part of the dynamic procedure, varied. This means that the dynamic procedure is able to distinguish between children in terms of the benefit shown during the dynamic procedure and, if the child is shown to have benefited, the type of help he/she has benefited from. This information is relevant within the scope of intervention planning, because it can help tailor the intervention to the needs of the child in terms of the type of instruction that is most helpful. Furthermore, the developmental psychologists evaluated the dynamic procedure positively. They particularly appreciated the practical usability and the extra information about the benefit shown, task behavior, communication abilities, and instruction needs, which could then be used in intervention planning. Their critical remarks mainly involved the extra time investment required, and the practicability of the scoring and interpretation procedure. Overall, the test and interview results supported the claim of enhanced clinical utility due to the dynamic procedure.

Viewed as a whole, the results of the studies that were part of the current thesis show that the accommodations that form the basis for the Low Motor/Vision and Low Verbal versions did not change the content of the Bayley-III-NL. This serves as justification for continued use of the standard norms in the accommodated versions. The results also show that the accommodations have added value for at least a subgroup of children in the target populations and that the dynamic version improves the usefulness of the Bayley-III-NL test results when it comes to designing an intervention.

7.2 Issues Encountered

A few issues were encountered during the research project. They are relevant with respect to each of the separate studies carried out, and provide an indication of the limitations and strengths of the studies. We will discuss the four issues that we deem to be the most important in this regard.

The first issue is related to the target group of young children with developmental disabilities. The characteristics of the children who belong to this target group are only roughly known. Consequently, it is not possible to know when a sample is representative for the target population. In addition, the target group is known to be very heterogeneous. Even children with the same diagnosis can vary greatly in their specific impairments and limitations. A heterogeneous target group can hinder the generalization of research results. Taking into account these limitations, our aim was for an optimal degree of representativeness for the sample, and thus generalizability of the research results, by including a relatively large and broad sample of children who were in need of a developmental assessment. We recruited children from a large number of different organizations, widely varying in terms of client groups, and located all over the Netherlands. The fact that we did manage to obtain a relatively large sample is due to this recruitment procedure. We were successful in asking developmental psychologists to refer children who were in need of a developmental assessment as part of the early intervention they received. The test results could then be used for the support of these children and for our research, simultaneously. This procedure served to lighten the workload of the developmental psychologists and resulted in a relatively large number of children being referred.

The second issue is related to the time limits that are basic to the Bayley-III-NL. Children with motor or visual impairments generally need more time to accomplish a test task as a consequence of their impairment. To take this into account in the Bayley-III-NL Low Motor/Vision version, one possibility was to lengthen the time limits for the speed items. This would, however, end up being rather arbitrary. It is difficult to make a well-founded argument for lengthening the time limits by a factor of two, for example. More importantly, there might still be children who continue to be unable to show their abilities despite the lengthened time limits. We therefore chose to completely eliminate time limits from the item instructions. The result might be that the elimination of the time limit would make an item easier to perform. This would especially influence the results on the

Cognition scale, which contains by far the largest number of speed items (19 out of the 91 items). The test results in the study on the Low Motor/Vision version, however, do not hint at any influence from the removal of time limits on the difficulty of the items.

The third issue relates to the fact that standardized developmental assessment instruments, including the Bayley-III, assume an ordinal development of skills and no large qualitative differences in development among children (Petermann & Macha, 2008). This assumption is related to the *similar sequence hypothesis* (Weisz & Zigler, 1979), which states that children with developmental disabilities develop in the same order, but at a slower rate, compared to typically developing children. This assumption is directly related to the structure of the Bayley-III, in which items of increasing difficulty (on average) are administered to the child, and the specific set of items that is administered depends on the responses of the child during the test administration. Obviously, if this assumption does not hold true, the risk is that the test results are invalid. An underestimation of developmental level can occur, for example, if a child fails five items in a row (which is the ceiling rule of the Bayley-III) and, consequently, items higher up in the scale are not administered, even though the child might well have been able to obtain a positive score on these “more difficult” items. In cases of functional impairment and/or severe developmental delay, especially, there is no consensus about whether the assumption of ordinal development actually holds true (Visser, Ruiter, Van der Meulen, Ruijssenaars, & Timmerman, 2012).

The fourth issue is related to our own research on the Low Verbal and dynamic versions of the Bayley-III-NL. With respect to the Low Verbal version, we were unable to compare the accommodated version directly to the standard version using double test administrations, as we had done with the Low Motor/Vision version. The reason for this was that the pilot study revealed that it is impossible to administer the standard version to a child with language impairment; this is because the test administrator unintentionally tends to apply (unstandardized) accommodations. We were, however, able to find indirect support for the claim that the Low Verbal accommodations did not change the content and difficulty of the test items by comparing the test results to those of a large group of typically developing children on the standard Bayley-III-NL. Furthermore, we were able to combine different sources of information and research methods by adding a more qualitatively oriented part to the study, which took the form of interviews.

This combination of different sources of information was also the strength of our study of the dynamic version. Given the limited scope of the current research project, it soon became evident that it would be impossible to obtain a firm quantitative foundation for the validity of test results obtained using the dynamic version. A firm quantitative foundation would have required test administrations to a very large group of children in order to obtain norms for the help-step scores. Furthermore, we would need to take outcome measures a few years from now in order to relate the difference score between the first and second test administrations to the outcome we now have. This would serve as an opportunity to validate the difference score in terms of whether it serves as a measure of learning potential. However, past research has proven that finding support for the predictive validity of developmental measures of young children is difficult, especially with respect to the first two years of life (Luttikhuizen dos Santos, De Kieviet, Königs, Van Elburg, & Oosterlaan, 2013; Greene, Patra, Silvestri, & Nelson, 2013; Lobo & Galloway, 2013; Petermann & Macha, 2008). It was therefore important for us to find an alternative way of studying the added value of the dynamic version. The combination of a description of the test results, along with an extensive description of the interview results, resulted in a clear picture of the added value of the dynamic version.

7.3 Implications and Suggestions for Early Intervention Practice

The current research project will result in the publication of the Bayley-III-NL Special Needs Addition in the Netherlands in 2014, which will consist of the Low Motor/Vision, the Low Verbal, and the dynamic versions. The availability of this instrument will expand the possibilities of making a fair developmental assessment of children with developmental disabilities and will improve the intervention orientation of the Bayley-III-NL. A fair developmental assessment is essential for tailoring the subsequent intervention to the needs and abilities of the child (Individuals with Disabilities Education Improvement Act, 2004a), which is of high clinical value, because a well-tailored intervention at an early age appears to be relatively effective (Shonkoff & Phillips, 2000).

The results of the study into the Low Motor/Vision version show that, on average, the accommodations lead to a higher score on the Cognition scale for children with motor and/or visual impairment, but not on the two Communication and the two Motor scales. We nevertheless recommend applying the accommodations to all five scales of the Bayley-III-NL, because this will probably result in a smoother test administration, even if the added value ends up not being reflected in all the test scores. Moreover, it is more convenient for the test administrator to use the accommodated test materials for all the scales, rather than switching to the standard materials after finishing the Cognition scale.

The studies into the Low Motor/Vision and the Low Verbal versions revealed that, although the majority of children do benefit from the accommodations, there are still children for whom the accommodations do not influence the test score or the smoothness of the test administration. Therefore, we recommend becoming conversant with the Special Needs Addition and then deciding whether or not to apply the accommodations for each child on an individual basis. The Special Needs Addition has been developed in such a way that it is easy to master for people who already know the standard Bayley-III-NL. The item instructions are integrated with those for the standard version, and the scoring forms contain extra directions for an accommodated test administration, including pictograms. Another recommendation is to keep in mind the option of applying only some of the accommodations. For a very small child with motor impairment, for example, it might be appropriate to support the elbows and eliminate the time limits, but not to use the enlarged materials, because these might be too big for that child.

When using the dynamic version, the test results should be interpreted with caution. The results with respect to the help-steps and learning potential should only be used as qualitative information that can be very valuable as a basis for intervention planning. This information can help tailor the intervention to the needs of the child in terms of the type of instruction that is most helpful. The results should not, however, be used for predicting future development because the predictive validity has not been examined. The dynamic version could be of added value in all situations, in which additional information for intervention planning is desired. However, in a situation in which the time available for assessment is limited, one solution might be to apply the dynamic procedure only to those children for whom additional qualitative information is expected to be specifically valuable. Examples of target groups, for whom the dynamic procedure could be specifically suitable, are children in the preliminary stage of school placement or from socially disadvantaged families.

7.4 Suggestions for Future Research

The current research project has improved the possibilities for making a fair developmental assessment of children with developmental disabilities in the Netherlands. Outside the Netherlands, there is no prospect yet in sight for a Special Needs Addition to the Bayley-III. It is, however, just as necessary elsewhere and so would prove just as useful in other countries, as it has in the Netherlands. Future work should therefore focus on developing Special Needs Additions to the Bayley-III in other languages and cultures as well. The current research project supports the claim of increased validity and unchanged content for the test, which in turn supports the use of the standard norms for the accommodated versions. If a Special Needs Addition in other languages is developed on the basis of the Dutch version, in such a way that only the language used is the difference, then we believe that the current study should provide sufficient support for its construct validity as well.

Another important goal for future research would be the development of an instrument for those target groups for whom no suitable standardized developmental assessment instrument is yet available. Examples of such target groups are children with deafness, blindness, or without any motor control in the hands and arms. For these groups, the accommodations to the Bayley-III-NL Special Needs Addition are not sufficient to obtain a fair assessment.

One more target group, for whom standardized developmental assessments are generally not suitable, are children and adults with Profound Intellectual and Multiple Learning Disabilities (PIMD), who have a developmental age equivalent of 2 years or less (Pawlyn & Carnaby, 2009). Child limitations, examiner limitations, as well as measurement limitations negatively influence the validity of psychological assessments in this target group (Burns, 2003). During the current research project, developmental psychologists did in fact indicate that there was a demand for an instrument that could fairly assess the developmental level of children and adults with PIMD. The Bayley-III-NL Low Motor/Vision version could be used for this purpose, but this should be studied in more detail for two reasons.

The first is that the assumption of ordinal development might not hold for people with PIMD (Van der Putten, Vlaskamp, Reynders, & Nakken, 2005). Research into Differential Item Functioning (DIF) of Bayley-III items in cases of PIMD could be one approach for studying this subject.

The second reason is related to the use of the Bayley-III for testing children who are older than 42 months of age, which is the upper limit of the Bayley-III age range. In the manual (Bayley, 2006a), the possibility of testing older children with the Bayley-III is mentioned. However, it is not possible to derive standard scores from the raw test scores, when a child is older than the upper age limit of the standardization sample of the test. This is not a problem when testing for research purposes, because the raw scores can be used in the analyses, just as we did in the current study. For interpreting individual test scores, however, the raw scores are not informative, and therefore developmental age equivalents need to be used as an alternative to the standard scores. Developmental age equivalents should be interpreted with caution, because small differences in raw scores can yield large differences in developmental age equivalents (American Educational Research Association [AERA], American Psychological Association [APA], & National Council on Measurement in Education [NCME], 1999; Evers & Resing, 2007; S. Johnson & Marlow, 2006). On the other hand, some authors suggest that using developmental quotient scores could be suitable in cases of severe developmental delay (Milne, McDonald, & Comino, 2012). Therefore, more research is necessary into the use of developmental age equivalents and the application of the Bayley-III in cases of a calendar age above 42 months.

A more general point for future research is related to the two issues identified in the literature review (Chapter 2) but that were beyond the focus of the current thesis. The first is the low reliability in cases of low levels of developmental functioning. This is a recurring problem in developmental assessment and could be solved by extending the number of items in the low range of the scales. The second issue is the long test duration of both the Bayley-III and the Special Needs Addition. The duration of testing of both the standard and accommodated versions of the Bayley-III can run to 90 minutes for all five scales, which is very long for young children, and can negatively influence the validity of the test administration. An important aim for future research into the Bayley-III should therefore be to find ways to shorten the test duration, while keeping the same level of accuracy.

References

Aiken, L. R., & Groth-Marnat, G. (2006). In K. Bowers (Ed.), *Psychological testing and assessment* (12th ed.). Boston: Pearson Education Group.

Alant, E., & Casey, M. (2005). Assessment concessions for learners with impairments. *South African Journal of Education, 25*, 185-189.

Albers, C. A., & Grieve, A. J. (2007). Review of "Bayley Scales of Infant and Toddler Development - Third Edition." *Journal of Psychoeducational Assessment, 25*, 180-190.

Allen, S. F. (2007). Assessing the development of young children in child care: A survey of formal assessment practices in one state. *Early Childhood Education Journal, 34*, 455-465. doi:10.1007/s10643-007-0153-z

American Academy of Pediatrics (2001). Developmental surveillance and screening in infants and young children. *Pediatrics, 108*, 192-196.

American Educational Research Association [AERA], American Psychological Association [APA], & National Council on Measurement in Education [NCME]. (1999). *Standards for educational and psychological testing*. Washington, D.C.: American Educational Research Association.

Americans with Disabilities Act, P.L. 101–336 (1990).

Bagnato, S. J., Macey, M., Salaway, J., & Lehman, C. (2007). *Research foundations of conventional tests and testing to ensure accurate and representative early intervention eligibility*. Pittsburgh, PA: Trace Center.

Bagnato, S. J., Neisworth, J. T., & Pretti-Frontczak, K. (2010). *LINKing authentic assessment and early childhood intervention: Best measures for best practice (2nd ed.)*. Baltimore, MD US: Paul H. Brookes.

Bain, S. K., & Allin, J. D. (2005). Review of "Stanford-Binet Intelligence Scales, Fifth Edition." *Journal of Psychoeducational Assessment, 23*, 87-95. doi:10.1177/073428290502300108

Bain, S. K., & Gray, R. (2008). Test reviews: Kaufman Assessment Battery for Children. *Journal of Psychoeducational Assessment, 26*, 92-101. doi:10.1177/0734282907300461

Barnett, A. L., Guzzetta, A., Mercuri, E., Henderson, S. E., Haataja, L., Cowan, F., et al. (2004). Can the Griffiths scales predict neuromotor and perceptual-motor impairment in term infants with neonatal encephalopathy? *Archives of Disease in Childhood, 89*, 637-643.

Batshaw Claire, E., Church, R. P., & Batshaw, M. L. (2007). Special education services. In M. L. Batshaw, L. Pellegrino & N. J. Roizen (Eds.), *Children with disabilities (6th ed.).* (pp. 523-538). Baltimore, MD: Paul H. Brookes.

Bayley, N. (1993). *Bayley Scales of Infant Development - Second Edition*. San Antonio, TX: The Psychological Corporation.

Bayley, N. (2006a). *Bayley Scales of Infant and Toddler Development - Third Edition*. San Antonio, TX: Harcourt Assessment.

Bayley, N. (2006b). *Bayley Scales of Infant and Toddler Development - Third Edition. Technical manual*. San Antonio, TX: Harcourt Assessment, Inc.

Becker, K. A. (2003). *History of the Stanford-Binet Intelligence Scales: Content and psychometrics (Stanford-Binet Intelligence Scales, Fifth Edition assessment service bulletin no. 1)*. Itasca, IL: Riverside.

Beran, T. N. (2007). Review of 'Differential Ability Scales' (2nd ed.). *Canadian Journal of School Psychology, 22*, 128-132. doi:10.1177/0829573507302967

Best, A. B., & Corn, A. L. (1993). The management of low vision in children: Report of the 1992 World Health Organization consultation. *Journal of Visual Impairment & Blindness, 87*, 307-309.

Blackwell, T. L. (2001). Test review: Woodcock, R. W., McGrew, K. S., & Mather, N. (2001). Woodcock-Johnson III test. Riverside Publishing Company. Itasca, IL. *Rehabilitation Counseling Bulletin, 44*, 232-235.

Bliss, S. L. (2007). Test reviews: Newborg, J. (2005). "Battelle Developmental Inventory – Second Edition." Itasca, IL: Riverside. *Journal of Psychoeducational Assessment, 25*, 409-415.

Bondurant-Utz, J. A. (2002). *Practical guide to assessing infants and preschoolers with special needs*. Upper Saddle River, NJ: Merrill Prentice Hall.

Bracken, B. A. (1987). Limitations of preschool instruments and standards for minimal levels of technical adequacy. *Journal of Psychoeducational Assessment, 5*, 313-326. doi:10.1177/073428298700500402

Bracken, B. A. (2000). Maximizing construct relevant assessment: The optimal preschool testing situation. In B. A. Bracken (Ed.), *The psychoeducational assessment of preschool children* (3rd ed., pp. 33-44). Boston: Allyn & Bacon.

Braden, J. P., & Elliot, S. N. (2003). Accommodations on the Stanford-Binet Intelligence Scales, Fifth Edition. *Stanford Binet Intelligence Scales, Fifth Edition, Assessment Bulleting Number 2*. Itasca, IL: Riverside.

Bradley-Johnson, S. (1994). *Psychoeducational assessment of students who are visually impaired or blind: Infancy through high school* (2nd ed.). Austin, TX: Pro-Ed.

Bradley-Johnson, S. (2001). Cognitive assessment for the youngest children: A critical review of tests. *Journal of Psychoeducational Assessment, 19*, 19-44. doi:10.1177/073428290101900102

Bradley-Johnson, S., & Durmusoglu, G. (2005). Evaluation of floors and item gradients for reading and math tests for young children. *Journal of Psychoeducational Assessment, 23*, 262-278.

Brambring, M. (2007). Divergent development of manual skills in children who are blind or sighted. *Journal of Visual Impairment & Blindness, 101*, 212-225.

Brigance, A. H. (2004). *Brigance Diagnostic Inventory of Early Development - Second Edition (IED-II)*. North Billerica, MA: Curriculum Associates.

Broer van Dijk - Van der Hulst, M.J., Laurent de Angulo, M. S., Peirens, G., De Ridder - Sluiter, J.G., Stam - Van den Doel, H.H., Van der Stege, H. A., & Steinbuch-Linstra, I. (2005). Ontwikkelingsproblemen en ontwikkelingsstoornissen [Developmental problems and developmental disorders]. In M. S. Laurent de Angulo, E. A. Brouwers-De Jong, J. F. M. Bijlsma-Schlösser, A. M. W. Bulk-Bunschoten, J. H. Pauwels & I. Steinbuch-Linstra (Eds.), *Ontwikkelingsonderzoek in de jeugdgezondheidszorg. Het Van Wiechen onderzoek. De Baecke-Fassaert Motoriektest [Developmental research in youth health care. The Van Wiechen study. The Baecke-Fassaert locomotion test]* (3rd ed. ed., pp. 92-108). Assen, The Netherlands: Koninklijke Van Gorcum.

Burns, C. W. (2003). Assessing the psychological and educational needs of children with moderate and severe mental retardation. In C. R. Reynolds, & R. W. Kamphaus (Eds.), *Handbook of psychological & educational assessment of children. Intelligence, aptitude, and achievement* (Second ed., pp. 671-684). New York: The Guilford Press.

Carlson, J. S., & Wiedl, K. H. (1992). Principles of dynamic assessment: The application of a specific model. *Learning and Individual Differences, 4.2*, 153-166.

Conlon, C. J. (2002). Early intervention. In M. L. Batshaw (Ed.), *Children with disabilities* (5th ed., pp. 579-588). Baltimore, MD: Paul H. Brookes.

Crais, E. R. (2009). Testing and beyond: Strategies and tools for evaluating and assessing infants and toddlers. *Language, Speech, and Hearing Services in Schools, 42*, 341-364.

De Leon, V., Bosa, C., Hugo, C., & Hutz, C. S. (2004). Psychometric properties of the PsychoEducational Profile Revised: PEP-R. *Avaliaçao Psicologica, 3*, 39-52.

Dutton, G. N., McKillop, E. C. A., & Saidkasimova, S. (2006). Visual problems as a result of brain damage in children. *British Journal of Opthalmology, 90*, 932-933.

Early Childhood Measurement and Evaluation Resource Centre. (2007). Tool review - Brigance Inventory of Early Development II. Retrieved July 27, 2010, from http://www.cup.ualberta.ca/index.php?option=com_docman&task=doc_details&gid=71

Elliott, C. D. (1996). *The British Ability Scales II*. Windsor, Berkshire: NFER-NELSON Publishing Company.

Elliott, C. D. (2007). *Differential Ability Scales - Second edition*. San Antonio, TX: Harcourt Assessment.

Emde, R. N., & Robinson, J. (2000). Guiding principles for a theory of early intervention: A developmental-psychoanalytic perspective. In J. P. Shonkoff & S. J. Meisels (Eds.), *Handbook of early childhood intervention* (2nd ed., pp. 160-178). Cambridge, United Kingdom: Cambridge University Press.

Evers, A., & Resing, W. C. M. (2007). Het drijfzand van didactische leeftijdsequivalenten. *Psycholoog*, 466-472.

Feuerstein, R., Klein, P. S., & Tannenbaum, A. J. (Eds.). (1991). *Mediated Learning Experience (MLE): Theoretical, psychosocial and learning implications.* Freund Publishing House.

Flanagan, D. P., & Alfonso, V. C. (1995). A critical review of the technical characteristics of new and recently revised intelligence tests for preschool children. *Journal of Psychoeducational Assessment, 13*, 66-90.

Flynn, J. R. (1984). The mean IQ of Americans: Massive gains 1932 to 1978. *Psychological Bulletin, 95*, 29-51. doi:10.1037/0033-2909.95.1.29

Freeman, L., & Miller, A. (2001). Norm-referenced, criterion-referenced, and dynamic assessment: What exactly is the point? *Educational Psychology in Practice, 17*, 3-16.

Friberg, J. C. (2010). Considerations for test selection: How do validity and reliability impact diagnostic decisions? *Child Language Teaching and Therapy, 26*, 77-92.

Fu, C., Hsieh, C., Tseng, M., Chen, Y., Huang, W., Wu, P., et al. (2010). Inter-rater reliability and smallest real difference of the Chinese PsychoEducational Profile - Third Edition for children with autism spectrum disorder. *Research in Autism Spectrum Disorders, 4*, 89-94. doi:10.1016/j.rasd.2009.09.002

Getman, G. N. (1993). *How to develop your childs intelligence*. Santa Ana, CA: Optometric Extension Program.

Glascoe, F. P. (2005). Screening for developmental and behavioral problems. *Mental Retardation in Developmental Disabilities Research Reviews, 11*, 173-179.

Gravetter, F. J., & Forzano, L. A. (2012). In Gravetter F. J., Forzano L. A. (Eds.), *Research methods for the behavioral sciences* (4th ed.). Belmont, CA: Wadsworth.

Greene, M. M., Patra, K., Silvestri, J. M., & Nelson, M. N. (2013). Re-evaluating preterm infants with the bayley-III: Patterns and predictors of change. *Research in Developmental Disabilities, 34*, 2107-2117. doi:10.1016/j.ridd.2013.04.001

Gregory, R. J. (2007). *Psychological testing: History, principles and applications* (5th ed.). Boston: Pearson Education.

Griffiths, R., & Huntley, M. (1996). *Griffiths mental development scales-revised: Birth to 2 years*. Oxford, United Kingdom: Hogrefe.

Grigorenko, E. L., & Sternberg, R. J. (1998). Dynamic testing. *Psychological Bulletin, 124*, 75-111. doi:10.1037/0033-2909.124.1.75

Groenveld, M. (1990). The dilemma of assessing the visually impaired child. *Developmental Medicine and Child Neurology, 32*, 1105-1113.

Guralnick, M. J., & Conlon, C. J. (2007). Early intervention. In M.L. Batshaw, L. Pellergrino, & N. J. Roizen (Ed.), *Children with disabilities* (6th ed., pp. 511-521). Baltimore, MD: Paul H. Brookes.

Hatton, D. D., Bailey, D. B., Burchinal, M. R., & Ferrell, K. A. (1997). Developmental growth curves of preschool children with vision impairments. *Child Development, 68*, 788-806.

Haywood, H. C., & Lidz, C. S. (2007). *Dynamic assessment in practice: Clinical and educational applications*. New York, NY: Cambridge University Press.

Hebbeler, K., Barton, L. R., & Mallik, S. (2008). Assessment and accountability for programs serving young children with disabilities. *Exceptionality, 16*, 48-63.

Heineman, K. R., & Hadders-Algra, M. (2008). Evaluation of neuromotor function in infancy - A systematic review of available methods. *Journal of Developmental and Behavioral Pediatrics, 29*, 315-323.

Hessels-Schlatter, C. (2002). A dynamic test to assess learning capacity in people with severe impairments. *American Journal on Mental Retardation, 107*, 340-351.

Hill, V. (2005). Through the past darkly: A review of the British Ability Scales Second Edition. *Child and Adolescent Mental Health, 10*, 87-98. doi:10.1111/j.1475-3588.2004.00123.x

Hoekstra, A. T., Jansen, G. G. H., Van der Meulen, B. F., Oenema-Mostert, C. E., & Ruijssenaars, A. J. J. M. (2010). Home intervention. Validating the item order of a developmental checklist. *Infants & Young Children, 23*, 303-311.

Individuals with Disabilities Education Improvement Act of 2004, pub. L. no. 108-446 §118 stat. 2647, (2004a).

Individuals With Disabilities Education Improvement Act of 2004, §300.43(c) stat. 15, (2004b).

Johnson, C. J., Beitchman, J. H., & Brownlie, E. B. (2010). Twenty-year follow-up of children with and without speech-language impairments: Family, educational, occupational, and quality of life outcomes. *American Journal of Speech-Language Pathology, 19*, 51-65.

Johnson, M. R., Wilhelm, C., Eisert, D., & Halperin-Phillips, D. (2001). Assessment of children with motor impairments. In R. J. Simeonsson, & S. L. Rosenthal (Eds.), *Psychological and developmental assessment: Children with disabilities and chronic conditions.* (pp. 205-224). New York, NY: Guilford Press.

Johnson, S., & Marlow, N. (2006). Developmental screen or developmental testing? *Early Human Development, 82*, 173-183. doi:10.1016/j.earlhumdev.2006.01.008

Kahn, R. J. (2000). Dynamic assessment of infants and toddlers. In C. S. Lidz, & J. G. Elliott (Eds.), *Advances in cognition and educational practice. Volume 6. Dynamic assessment: Prevailing models and applications* (pp. 325-373). New York: Elsevier.

Kahn, J. V. (1976). Utility of the Uzgiris and Hunt scales of sensorimotor development with severely and profoundly retarded children. *American Journal of Mental Deficiency, 80*, 663-665.

Kaufman, A. S., & Kaufman, N. L. (2004). *Kaufman Assessment Battery for Children, Second edition*. Circle Pines, MN: AGS.

Kaufman, A. S., Lichtenberger, E. O., Fletcher-Janzen, E., & Kaufman, N. L. (2005). *Essentials of KABC-II assessment*. Hoboken, NJ: Wiley.

Kikas, E., & Haïdkind, P. (2003). Developing an Estonian version of the PsychoEducational Profile Revised (PEP-R). *Journal of Autism and Developmental Disorders, 33*, 217.

Klinger, L. G., O'Kelley, S. E., & Mussey, J. L. (2009). Assessment of intellectual functioning in autism spectrum disorders. In S. Goldstein, J. A. Naglieri, & S. Ozonoff (Eds.), *Assessment of autism spectrum disorders* (pp. 209-252). New York: The Guilford Press.

Laros, J. A., Tellegen, P. J., de Jesus, G. R., & Karino, C. A. (2011). *SON-R 2½-7[a]: Teste não-verbal de inteligência. manual com normatização e validação Brasileira*. Porto Alegre: Casa do Psicólogo.

Lauchlan, F., & Elliot, J. (2001). The psychological assessment of learning potential. *British Journal of Educational Psychology, 71*, 647-665.

Lidz, C. S. (Ed.). (1987). *Dynamic assessment: An interactional approach to evaluating learning potential.* Guilford Press.

Lobo, M. A., & Galloway, J. C. (2013). Assessment and stability of early learning abilities in preterm and full-term infants across the first two years of life. *Research in Developmental Disabilities, 34*, 1721-1730. doi:10.1016/j.ridd.2013.02.010

Lorenzo-Seva, U., & Ferrando, P. J. (2006). FACTOR: A computer program to fit the exploratory factor analysis model. *Behavioral Research Methods, Instruments and Computers, 38*, 88-91.

Luiz, D. M., Barnard, A., Knoesen, N., Kotras, N., Horrocks, S., McAlinden, P., et al. (2006). *Griffiths Mental Development Scales – Extended Revised. Two to eight years. Administration manual*. Oxford, United Kingdom: Hogrefe.

Luiz, D. M., Foxcroft, C. D., & Tukulu, A. N. (2004). The Denver II Scales and the Griffiths Scales of Mental Development: A correlational study. *Journal of Child and Adolescent Mental Health, 16*, 77-81.

Luttikhuizen dos Santos, E.S., De Kieviet, J. F., Königs, M., Van Elburg, R. M., & Oosterlaan, J. (2013). Predictive value of the Bayley Scales of Infant Development on development of very preterm/very low birth weight children: A meta-analysis. *Early Human Development, 89*, 487-496. doi:10.1016/j.earlhumdev.2013.03.008

Macha, T., Mayer, H., Petermann, F., Petermann, U., & Waldeck, E. (2007). Einsatz von entwicklungstests in der rehabilitation von kindern mit bewegungsstörungen [Developmental assessment in rehabilitation of children with cerebral palsy]. In Deutsche Rentenversicherung (Ed.), *16. Rehabilitationswissenschaftliches kolloquium [Report of the 16th rehabilitation science colloquium]* (pp. 567-568). Munich, Germany: DRV.

Macha, T., & Petermann, F. (2008). The ET 6–6. A method for developmental assessment for German-speaking countries. *Zeitschrift Für Psychologie / Journal of Psychology, 216*, 154-160.

Marshall, S., McGoey, K. E., & Moschos, S. (2011). Test review: C. D. Elliott "Differential Ability Scales - Second Edition." San Antonio, TX - Harcourt Assessment, 2007. *Journal of Psychoeducational Assessment, 29*, 89-93.

Mashburn, A. J., & Myers, S. S. (2010). Advancing research on children with speech-language impairment: An introduction to the early childhood longitudinal Study - Kindergarten cohort. *Language, Speech, and Hearing Services in Schools, 41*, 61-69.

Mayes, S. D. (1999). Mayes motor-free compilation (MMFC) for assessing mental ability in children with physical impairments. *International Journal of Disability, Development and Education, 46*, 475-485.

McCabe, P. C. (2005). Social and behavioral correlates of preschoolers with specific language impairment. *Psychology in the Schools, 42*, 373-387.

McLaren, C., Edwards, G., Ruddick, S., Zabjek, K., & McKeever, P. (2011). Kindergarten kids in motion: Rethinking inclusive classrooms for optimal learning. *Educational & Child Psychology, 28*, 100-113.

Meijer, R. R., & Baneke, J. J. (2004). Analyzing psychopathology items: A case for nonparametric item response theory modeling. *Psychological Methods, 9*, 354-368. doi:10.1037/1082-989X.9.3.354

Meisels, S. J. (1994). Designing meaningful measurements for early childhood. In B. L. Mallory & R. S. New (Eds.), *Diversity in early childhood education: A call for more inclusive theory, practice, and policy* (pp. 205-225). New York: Teachers College Press.

Meisels, S. J., & Atkins-Burnett, S. (2000). The elements of early childhood assessment. In J. P. Shonkoff, & S. J. Meisels (Eds.), *Handbook of early childhood intervention* (2nd ed., pp. 231-257). Cambridge, United Kingdom: Cambridge University Press.

Melchers, P., & Preuβ, U. (2009). *Kaufman Assessment Battery for Children.* 8. unveränderte auflage. Frankfurt, Germany: Pearson Assessment.

Miller, J. C., & Skillman, G. D. (2003). Assessors' satisfaction with measures of cognitive ability applied to persons with visual impairments. *Journal of Visual Impairment & Blindness, 97*, 769-774.

Milne, S., McDonald, J., & Comino, E. J. (2012). The use of the Bayley Scales of Infant and Toddler Development III with clinical populations: A preliminary exploration. *Physical & Occupational Therapy in Pediatrics, 32*, 24-33. doi:10.3109/01942638.2011.592572

Mokken, R. J. (1971). *A theory and procedure of scale analysis*. The Hague, The Netherlands: Mouton.

Molenaar, I. W., & Sijtsma, K. (2000). *MSP5 for windows, a program for Mokken scale analysis for polytomous items*. Groningen, The Netherlands: iec ProGAMMA.

Naglieri, J. A., & Chambers, K. M. (2009). Psychometric issues and current scales for assessing autism spectrum disorders. In S. Goldstein, J. A. Naglieri, & S. Ozonoff (Eds.), *Assessment of autism spectrum disorders* (pp. 55-90). New York: Guilford Press.

National Association for the Education of Young Children & National Association of Early Childhood Specialists in State Departments of Education (2003). *Early childhood curriculum, assessment, and program evaluation. Building an effective, accountable system in programs for children birth through age 8.* Retrieved November 18, 2011, from http://www.naeyc.org/files/naeyc/file/positions/pscape.pdf

Neisworth, J. T., & Bagnato, S. J. (1992). The case against intelligence testing in early intervention. *Topics in Early Childhood Special Education, 12*, 1-20.

Neisworth, J. T., & Bagnato, S. J. (2004). The mismeasure of young children: The authentic assessment alternative. *Infants and Young Children, 17*, 198-212.

Neutel, R. J., Van der Meulen, B. F., & Lutje Spelberg, H. C. (1996). *GOS 2½-4½. Groningse ontwikkelingsschalen. Handleiding* [Groningen developmental scales. Manual]. Lisse, The Netherlands: Swets & Zeitlinger.

Newborg, J. (2005). *Battelle Developmental Inventory* (2nd ed.). Itasca, IL: Riverside.

O'Connell, R. (1994). Proving integration works: How effective is the integration of students with PMLDs into the mainstream of an SLD school in increasing opportunities for social interaction? In J. Ware (Ed.), *Educating children with profound and multiple learning difficulties* (pp. 85-98). London, United Kingdom: David Fulton.

Orelove, F. P., & Sobsey, D. (1996). *Educating children with multiple disabilities: A transdisciplinary approach*. Baltimore, MD: Brooks.

Osterlind, S. J., & Everson, H. T. (2009). *Differential item functioning* (2nd ed.). Newbury Park, CA: Sage.

Oxford Dictionaries Online (2012a). Practicable. Retrieved, 2012, from http://oxforddictionaries.com/definition/practicable

Oxford Dictionaries Online (2012b). Suitable. Retrieved, 2012, from http://oxforddictionaries.com/definition/suitable

Pagliano, P. (1998). Students with vision impairment. *Educating Children with Special Needs,* 383-416.

Pameijer, N. K., & Van Beukering, J. T. E. (2004). *Handelingsgerichte diagnostiek: Een praktijkmodel voor diagnostiek en advisering bij onderwijsleerproblemen* [Needs based assessment: A practice-based model for diagnosis and counseling in cases of educational learning problems]. Acco.

Parveva, T., De Coster, I., & Noorani, S. (2009). *National testing of pupils in Europe: Objectives, organisation and use of results*. Brussels, Belgium: Education, Audiovisual and Culture Executive Agency.

Pawlyn, J., & Carnaby, S. (Eds.). (2009). *Profound intellectual and multiple disabilities: Nursing complex needs*. Oxford, United Kingdom: Blackwell Publishing.

Petermann, F., & Macha, T. (2008). Developmental assessment. A general framework. *Journal of Psychology, 216*, 127-134.

Petermann, F., Stein, I. A., & Macha, T. (2006). *Entwicklungsdiagnostik mit dem ET 6–6 [Developmental assessment using the ET 6–6]* (3rd ed.). Frankfurt, Germany: Harcourt Test Services.

Pijl, S. J., & Hamstra, D. (2005). Assessing pupil development and education in an inclusive setting. *International Journal of Inclusive Education, 9*, 181-192. doi:http://dx.doi.org/10.1080/1360311042000331192

Pro-Ed, Inc. (2011). DAYC: Developmental assessment of young children. Retrieved August 12, 2011, from http://www.proedinc.com/customer/productView.aspx?ID=1536

Provence, S., Erikson, J., Vater, S., & Palmeri, S. (1995). *Infant - toddler developmental assessment: IDA*. Chicago: Riverside.

Rathvon, N. (2004). *Early reading assessment: A practitioner's handbook*. New York: Guilford Press.

Rémillard, G. (2008). *Reconciling economics and health*. Montreal, Quebec: Décision Média.

Resing, W. C. M. (2006). Using children's 'ability to learn' in diagnosis and assessment. *Educational and Child Psychology, 23*, 6-11.

Ricken, G., Fritz, A., Schuck, K. D., & Preuß, U. (2007). *Hannover-Wechsler-Intelligenztest für das VorschulAlter – III (HAWIVA –III)* [Hannover-Wechsler intelligence test for preschoolers – III]. Göttingen, Germany: Hogrefe.

Rimrodt, S. L., & Johnston, M. V. (2009). Neuronal plasticity and developmental disability. In M. Shevell (Ed.), *Neurodevelopmental disabilities: Clinical and scientific foundations* (pp. 225-240). London, England: MacKeith Press.

Roid, G. H. (2003). *Stanford-Binet Intelligence Scales, Fifth Edition*. Itasca, IL: Riverside.

Roid, G. H., & Miller, L. J. (1997). *Leiter International Performance Scale - Revised*. Wood Dale, IL: Stoelting.

Rolfe, S. A. (1994). Does assessment of cognitive functioning in infancy hold the key to early detection of developmental disabilities? A review of research. *Australia and New Zealand Journal of Developmental Disabilities, 19*, 61-72.

Romanczyk, R. G., Gillis, J. M., Noyes-Grosser, D. M., Holland, J. P., Holland, C. L., & Lyons, D. (2005). Clinical clues, developmental milestones, and early identification/assessment of children with disabilities: Practical applications and conceptual considerations. *Infants & Young Children, 18*, 212-221.

Ruiter, S. A. J. (2007). *The BSID-II-NL for assessing children with specific impairments*. Groningen, The Netherlands: Stichting Kinderstudies.

Ruiter, S. A. J., Hoekstra, A. T., Van der Meulen, B. F., Lutje Spelberg, H. C., & Nakken, H. (2006). De niet-verbale versie van de BSID-II-NL. *Tijdschrift Voor Orthopedagogiek, 45*, 406-413.

Ruiter, S. A. J., Nakken, H., Janssen, M., Van der Meulen, B. F., & Looijestijn, P. (2011). Adaptive assessment of young children with visual impairment. *British Journal of Visual Impairment, 29*, 93-112.

Ruiter, S. A. J., Nakken, H., Van der Meulen, B. F., & Lunenborg, C. B. (2010). Low motor assessment: A comparative pilot study with young children with and without motor impairment. *Journal of Developmental and Physical Disabilities, 22*, 33-46. doi:10.1007/s10882-009-9165-5

Rydz, D., Shevell, M. I., Majnemer, A., & Oskoui, M. (2005). Developmental screening. *Journal of Child Neurology, 20*, 4-21. doi:10.1177/08830738050200010201

Sameroff, A. J. (Ed.). (2009). *The transactional model of development; How children and contexts shape each other*. Washington, DC: American Psychological Association.

Sattler, J. M. (2001). *Assessment of children: Cognitive applications* (4th ed.). La Mesa, CA: Author.

Sattler, J. M. (2008). *Assessment of children: Cognitive foundations* (5th ed.). La Mesa, CA: Author.

Schopler, E., Lansing, M. D., Reichler, R. J., & Marcus, L. M. (2005). *PsychoEducational Profile - Third Edition (PEP-3)*. Austin, TX: Pro-Ed.

Shonkoff, J. P. (2007). The science of early childhood development. Closing the gap between what we know and what we do. Retrieved March, 2010, from http://www.speaker.gov/pdf/Shonkoff.pdf

Shonkoff, J. P., & Phillips, D. A. (Eds.). (2000). *From neurons to neighborhoods: The science of early childhood development*. Washington, DC: National Academy Press.

Sijtsma, K., & Molenaar, I. W. (2002). *Introduction to nonparametric item response theory. volume 5*. Thousand Oaks: Sage Publications.

Skovgaard, A. M., Houmann, T., Landorph, S. L., & Christiansen, E. (2004). Assessment and classification of psychopathology in epidemiological research of children 0-3 years of age: A review of the literature. *European Child & Adolescent Psychiatry, 13*, 337-346.

Snow, C. E., & Van Hemel, S. B. (Eds.). (2008). *Early childhood assessment: Why, what, and how*. Washington DC: National Academy Press.

Sparrow, S. S., & Davis, S. M. (2000). Recent advances in the assessment of intelligence and cognition. *Journal of Child Psychology and Psychiatry, 41*, 117-131.

Steenis, L. J. P., Verhoeven, J. C. T., & Van Baar, A. L. (2012). The Bayley III: The instrument for early detection of developmental delay. In A.M. Columbus (Ed.), *Advances in Psychology Research* (pp. 133-141). Volume 92. Hauppauge, NY: Nova Science Publishers.

Steerneman, P., Muris, P., Merckelbach, H., & Willems, H. (1997). Assessment of development and abnormal behavior in children with pervasive disorders: Evidence for the reliability and validity of the revised PsychoEducational Profile. *Journal of Autism and Developmental Disorders, 27*, 177-185.

Sternberg, R. J., & Grigorenko, E. L. (2001). All testing is dynamic testing. *Issues in Education, 7*, 137-170.

Tellegen, P. J., Laros, J. A., & Heider, D. (2008). *SON-R 2½-7, neverbalni inteligencni test: Prirucka.* [SON-R 2½-7, non-verbal test of intelligence: Manual]. Praha, Czech Republic: Testcentrum-Hogrefe.

Tellegen, P. J., Laros, J. A., & Kiat, M. (2009). *SON-R 2½-7. Test non-verbal d'intelligence. Adaption Française. Manuel.* [SON-R-2½-7. Non-verbal test of intelligence, French Adaptation. Manual]. Paris, France: Editions Hogrefe France.

Tellegen, P. J., Laros, J. A., Kopcanova, D., Farkasova, E., & Dockal, V. (2009). *SON-R 2½-7, neverbalny test inteligencia. Prve Slovenske vydanie. Prirucka. [SON-R 2½-7, non-verbal test of intelligence: Slovak normatization. Manual].* Praha, Czech Republic: Vydalo Hogrefe – Testcentrum.

Tellegen, P. J., Laros, J. A., & Petermann, F. (2007). *SON-R 2½-7 non-verbaler intelligenztest. Testmanual mit Deutscher normierung und validierung*. [SON-R-2½-7 Non-verbal test of intelligence. Test manual with German standardization and validation]. Göttingen, Germany: Hogrefe Verlag.

Tellegen, P. J., & Laros, J. A. (2011). Tests & test-research. SON-R 2½-7 niet-verbale intelligentie test. [Tests & test-research. SON-R 2½-7 non-verbal intelligence test]. Retrieved June/11, 2013, from http://www.testresearch.nl/sonr/sonr2.html

Tellegen, P. J., Winkel, M., Wijnberg-Williams, B. J., & Laros, J. A. (1998a). *Snijders-Oomen Niet-verbale intelligentietest SON-R 2½-7. Handleiding. Tweede gereviseerde editie*. [Snijders-Oomen Non-verbal test of intelligence. SON-R 2½-7. Manual. Second revised edition]. Göttingen, Germany: Hogrefe Verlag.

Tellegen, P. J., Winkel, M., Wijnberg-Williams, B. J., & Laros, J. A. (1998b). *Snijders-Oomen Nonverbal intelligence test. SON -R 2½-7 manual and research report*. Lisse, The Netherlands: Swets & Zeitlinger.

Teplin, S. W. (1995). Visual impairment in infants and young children. *Infants and Young Children, 8*, 18-51.

Thal, D. J. (1991). Language and cognition in normal and late-talking toddlers. *Topics in Language Disorders, 11*, 33-42.

Thurlow, M. L., Elliot, J. L., & Ysseldyke, J. E. (2003). *Testing students with disabilities: Practical strategies for complying with district and state requirements* (2nd ed.). Thousand Oaks, CA: Corwin Press.

Tusing, M. E., Maricle, D. E., & Ford, L. (2003). Assessment with the Woodcock-Johnson III and young children. In F. Schrank, & D. P. Flanagan (Eds.), *WJ III clinical use and interpretation: Scientist and practitioner perspectives* (pp. 243-283). San Diego, CA: Academic Press.

Tzuriel, D. (2000). Dynamic assessment of young children: Educational and intervention perspectives. *Educational Psychology Review, 12*, 385-435.

Tzuriel, D. (2001). *Dynamic assessment of young children*. New York: Kluwer Academic/Plenum Publishers.

Van der Linden, W. J. (2008). Some new developments in adaptive testing technology. *Journal of Psychology, 216*, 3-11.

Van der Meulen, B. F., Ruiter, S. A. J., Spelberg, H. C. L., & Smrkovsky, M. (2002). *Bayley Scales of Infant Development - Second edition - Nederlandse versie*. Amsterdam, The Netherlands: Harcourt.

Van der Putten, A., Vlaskamp, C., Reynders, K., & Nakken, H. (2005). Movement skill assessment in children with profound multiple disabilities: A psychometric analysis of the top down motor milestone test. *Clinical Rehabilitation, 19*, 635-643.

Veenstra, B., Van Geert, P. L. C., & Van der Meulen, B. F. (2010). Computer versus human-based support: Effect on computer game performances in (in)effective learning pre-schoolers. *Educational & Child Psychology, 27*, 56-72.

Villa, S., Micheli, E., Villa, L., Pastore, V., Crippa, A., & Molteni, M. (2010). Further empirical data on the PsychoEducational Profile Revised (PEP-R): Reliability and validation with the Vineland Adaptive Behavior Scales. *Journal of Autism and Developmental Disorders, 40*, 334-341.

Vink, M. T. (1994). Vasthouden, loslaten, zelf doen. De vroege bewegingsontwikkeling van blinde kinderen. [Holding, letting go, do it yourself. The early movement development of blind children]. *Bewegen & Hulpverlening, 11*, 229-247.

Visser, L., Ruiter, S. A. J., Van der Meulen, B. F., Ruijssenaars, A. J. J. M., & Timmerman, M. E. (2012). A review of standardized developmental assessment instruments for young children and their applicability for children with special needs. *Journal of Cognitive Education and Psychology, 11*, 102-127. doi: 10.1891/1945-8959.11.2.102

Visser, L., Ruiter S.A.J., Van der Meulen, B.F., Ruijssenaars, A.J.J.M., Timmerman, M.E. (2013a). *Dynamic assessment with the Bayley-III among young children with developmental disabilities.* Manuscript submitted for publication.

Visser, L., Ruiter S.A.J., Van der Meulen, B.F., Ruijssenaars, A.J.J.M., Timmerman, M.E. (2013b). *Low verbal assessment with the Bayley-III*. Manuscript submitted for publication.

Visser, L., Ruiter, S. A. J., Van der Meulen, B. F., Ruijssenaars, A. J. J. M., & Timmerman, M. E. (2013c). Validity and suitability of the Bayley-III Low Motor/Vision version: A comparative study among young children with and without motor and/or visual impairments. *Research in Developmental Disabilities, 34*, 3736-3745. doi:10.1016/j.ridd.2013.07.027

Visser, L., Ruiter, S. A. J., Van der Meulen, B. F., Ruijssenaars, A. J. J. M., & Timmerman, M. E. (in press). Accommodating the Bayley-III with regard to motor and/or visual impairment: A comparative pilot study. *Pediatric Physical Therapy.*

Voress, J. K., & Maddox, T. (1998). *Developmental assessment of young children: Examiner's manual*. Austin, TX: Pro-Ed.

Vygotsky, L. S. (1978). *Mind in society: The development of higher psychological processes*. London, United Kingdom: Harvard University Press.

Ward, K. E., Rothlisberg, B. A., McIntosch, D. E., & Bradley, M. H. (2011). Assessing the SB-V factor structure in a sample of preschool children. *Psychology in the Schools, 48*, 454-463.

Wechsler, D. (2002a). *Wechsler Preschool and Primary Scale of Intelligence,Tthird edition (WPPSI-III).* San Antonio, TX: Pearson.

Wechsler, D. (2002b). *WPPSI-III technical and interpretive manual*. San Antonio, TX: The Psychological Corporation.

Wechsler, D. (2003a). *Wechsler Preschool and Primary Scale of Intelligence – Third edition (WPPSI-III UK).* Oxford, United Kingdom: The Psychological Corporation.

Wechsler, D. (2003b). *The Wechsler Preschool and Primary Scale of Intelligence Scale – Third edition, Canadian*. San Antonio, TX: The Psychological Corporation.

Wechsler, D. (2004a). *WPPSI-III: Échelle D'Intelligence de Wechsler pour la période pré-scolaire et primaire, Troisième édition* [Wechsler Preschool and Primary Scale of Intelligence, Third edition. WPPSI-III]. Paris, France: ECPA.

Wechsler, D. (2004b). *Wechsler Preschool & Primary Scale of Intelligence – Australian standardisation (WPPSI-III Australian)*. Sydney, Australia: Pearson Clinical and Talent Assessment Australia and New Zealand.

Wechsler, D. (2005). *WPPSI-III. Wechsler Preschool and Primary Scale of Intelligence. Swedish version.* Stockholm, Sweden: Psykologiförlaget AB.

Wechsler, D. (2008a). *Wechsler Preschool and Primary Scale of Intelligence-III (WPPSI-III). Norwegian version*. Stockholm, Sweden: Pearson Assessment.

Wechsler, D. (2008b). *WPPSI-III. Wechsler Preschool and Primary Scale of Intelligence - Third edition. Italian adaptation*. Firenze, Italy: Giunti OS.

Wechsler, D. (2009a). *Wechsler Preschool and Primary Scale of Intelligence – Third edition (WPPSI-III).* Helsinki, Finland: Psykologien Kustannus.

Wechsler, D. (2009b). *WPPSI-III, Escala de Intelligencia de Wechsler Para Preescolar y Primaria – III* [WPPSI-III: Wechsler Preschool and Primary Scale of Intelligence, Third edition. Spanish adaptation]. Madrid, Spain: TEA Ediciones.

Wechsler, D. (2009c). *WPPSI-III-NL. Nederlandstalige bewerking (Nederlandse bewerking van J. Hendriksen en P. Hurks).* Amsterdam, The Netherlands: Pearson Assessment and Information.

Wechsler, S. M., Nunes, C. S., Schelini, P. W., Pasian, S. R., Homsi, S. V., Moretti, L., et al. (2010). Brazilian adaptation of the Woodcock-Johnson III cognitive tests. *School Psychology International, 31*, 409-421.

Weisz, J. R., & Zigler, E. (1979). Cognitive development in retarded and nonretarded persons: Piagetian tests of the similar sequence hypothesis. *Psychological Bulletin, 86*, 831-851. doi:10.1037/0033-2909.86.4.831

Wever, G., De Klerk, A., & Van der Loos, P. (2006). *Kinderen met een beperking en kinderopvang; onderzoek naar de belemmeringen om arbeid en zorg te combineren voor ouders van een kind met een beperking of chronische ziekte.* [Children with an impairment and child care: research into the obstacles to combine work and care for parents of a child with an impairment or chronic disease]. (Unpublished report No. 3110876460/060518Do01). Rotterdam, The Netherlands: Deloitte Consultancy. Retrieved from http://www.samennaardekinderopvang.nl/nieuws/00001/Eindrapport_Deloitte_Kinderen_met_een_beperking_en_kinderopvang.pdf/

Whinnery, K. W., & Whinnery, S. B. (2007). MOVE: Systematic programming for early motor intervention. *Infants & Young Children, 20*, 102-108.

Winkel, M. (1999). *Snijders-Oomen Niet-verbale intelligentietest. Constructie, plaatsbepaling en validering van deze algemene niet-verbale intelligentietest voor jonge kinderen.* Groningen, The Netherlands: Stichting Drukkerij C. Regenboog.

Woodcock, R. W., McGrew, K. S., & Mathner, N. (2001a). *Woodcock-Johnson III: Tests of achievement*. Itasca, IL: Riverside.

Woodcock, R. W., McGrew, K. S., & Mather, N. (2001b). *Woodcock-Johnson III: Tests of cognitive abilities*. Itasca, IL: Riverside.

World Health Organization. (1992). *International statistical Classification of Diseases and related health problems, Tenth revision (ICD-10).* Geneva, Switzerland: World Health Organization.

World Health Organization. (2007). *International Classification of Functioning, Disability, and Health - Children and youth*. Geneva, Switzerland: World Health Organization (WHO).

World Health Organization [WHO]. (1992). *The ICD-10 Classification of mental and behavioural disorders* (10th ed.). Geneva, Switzerland: World Health Organization.

Ziviani, J., Darlington, Y., Feeney, R., Rodger, S., & Watter, P. (2013). Service delivery complexities. early intervention for children with physical disabilities. *Infants & Young Children, 26*, 147-163.

Summary

Development of children in the first few years of life usually happens relatively spontaneously. In a small percentage of children, however, early development is not trouble-free. Minor developmental delay can occur, which the child does or does not overcome at a later age. More serious problems can also occur, such as impairments that are caused by problems during the pregnancy or birth (e.g., cerebral palsy) or a genetic abnormality (e.g., Down syndrome). Sometimes the cause is not clear, as is often the case with general developmental delay or with a pervasive developmental disorder.

In cases of developmental problems, early recognition is very important, because intervention appears to be most successful when it is applied starting at an early age. This is due to the fact that children develop very quickly in the first years of life. By stimulating development at this early stage, the consequences of an impairment can be kept to a minimum. The aim of such early intervention is optimal development for the child, which is different for every child and depends on the degree of impairment, among other things. If parents or professionals suspect developmental problems, an assessment takes place to identify the problem. The administration of a standardized developmental assessment instrument is an important part of such an assessment. It can be used to identify the level of development of a child by comparing the test results to those of a large group of children of the same age. Different areas of development are generally part of this assessment, such as cognition, language, motor skills, and social-emotional development.

Standardized developmental assessment instruments are developed on the basis of test administrations with a large group of children without developmental problems. The test results of those typically developing children are used to develop the test's norms. Norms provide information about the distribution of test scores for typically developing children in particular age ranges. However, the nature of testing children with special needs is often not taken into account during the development of an instrument. As a consequence, many standardized developmental assessment instruments cannot adequately assess the development of children with special needs. An adequate assessment means that the assessment is both reliable (i.e., accurate) and valid, which means that it should assess what is meant to be assessed. To give an example: If cognitive development is the target of an

assessment, you do not want to measure the motor skills that are necessary to grasp the test materials. You want to measure the understanding that the child has of the task. However, many instruments for assessing cognitive development contain jigsaws or other test materials, whose handling relies too heavily on fine motor skills. If an impairment limits a child's ability to accomplish the test tasks, the child is not able to show his or her abilities. This may lead to an underestimation of the level of cognitive development and, consequently, inadequate support. To prevent this, it is important to ensure that test materials are just as suitable for children with impairments.

Another frequently reported shortcoming of standardized developmental assessment instruments is that they do not yield sufficient information that can be used as a basis for an intervention plan. Information about support needs, for example, is necessary in order to be able to gear the intervention to the child's developmental level and needs.

The aim of the research that is described in the current thesis is to develop a standardized developmental assessment instrument that is suitable for young children with an impairment. The Bayley-III-NL serves as the basis for this research. This is the Dutch version of the Bayley-III, which is a widely used and researched instrument for developmental assessment of children of up to 3½ years of age, and consists of the scales of Cognition, Receptive Communication, Expressive Communication, Fine Motor development, and Gross Motor development.

Chapter 1 of this thesis contains the introduction to the topic and a discussion of the relevance for early intervention practice. The research questions are described and the chapter concludes with an overview of the thesis.

In **Chapter 2**, we describe a literature review into available instruments for developmental assessment of children aged 0-4. The chapter starts with an outline of existing problems with respect to assessing children with special needs, followed by an overview of 18 instruments for developmental assessment, which we found during the literature search. We describe these instruments and discuss their suitability for assessing children with special needs. Our conclusion is that the range of available instruments is limited, especially for children below the age of 2.

Chapter 3 describes a pilot study on the Low Motor/Vision version of the Bayley-III-NL, which we developed to increase the suitability for children with motor and/or visual impairment. To this end, we applied accommodations to the standard Bayley-III-NL, which are adaptations that decrease the influence of an impairment on the test results, without changing what the instrument aims to measure. In the pilot study, 19 children with a motor and/or visual impairment were tested with both the Low Motor/Vision and the standard version of the Bayley-III-NL. The test results showed that some of the children did benefit from the accommodations, while others did not. The test administrators filled in an evaluation form for each child concerning the suitability of the accommodated instrument. They evaluated the accommodations as practicable and advantageous for a majority of the children. The conclusion is that the Bayley-III-NL Low Motor/Vision seems to adequately assess the development of children with motor and/or visual impairment but that research on a larger scale is necessary.

Chapter 4 describes a large-scale study on the Low Motor/Vision version. The sample consisted of 41 children without impairment and 63 children with a motor and/or visual impairment. All children were tested with both the standard and Low Motor/Vision versions, in counterbalanced order. As was the case in the pilot study, the test administrators filled in an evaluation form. The results show that the children without impairment obtained equal scores on the Low Motor/Vision and standard version of the Bayley-III-NL. This supports the assumption that the accommodations do not change the content and difficulty of the Bayley-III-NL items, as was our intent. With respect to the Cognition scale, the children with impairment obtained a higher score on the Low Motor/Vision version than on the standard version, on average. This means that the accommodations enhanced the children's opportunity to show their skills during the assessment. We did not find the same difference in test scores between the two versions for the Communication and Motor scales. The test administrators indicated that almost all of the children with impairment had been able to show their abilities during the test administration and that the accommodations were beneficial in a majority of cases. For a small number of children the Bayley-III-NL Low Motor/Vision version appeared unsuitable, because the child's impairment was too severe and the child could still not show his or her abilities on the test.

In **Chapter 5** we describe a study on the Low Verbal version of the Bayley-III-NL, which is designed for children with a language impairment and consists of an accommodated Cognition scale, and non-accommodated Communication and Motor scales. We tested 69 children with a language impairment using the Bayley-III-NL Low Verbal and compared these test results to those of a large sample of children without any impairment, tested on the standard version, by using Item Response Analysis (IRT). IRT is a statistical method that can be used to study the characteristics of test items. The results show that the items of the Bayley-III-NL Low Verbal, administered to children with a language impairment, have characteristics similar to items of the standard version administered to children without an impairment. The results of an evaluation form and interviews with developmental psychologists working in early intervention practice show that the Bayley-III-NL Low Verbal was suitable for the children in the sample and more suitable than the standard version in a majority of cases. In addition, the developmental psychologists indicated that they found the Low Verbal version also suitable for children with general developmental delay. The conclusion is that The Bayley-III-NL Low Verbal can validly assess the development of children with a language impairment.

In **Chapter 6** we describe a study into the dynamic version of the Bayley-III-NL. We added a dynamic procedure to the Cognition scale with the aim of enhancing the usefulness of the test results in setting up an intervention plan. The dynamic procedure starts with the administration of the standard Cognition scale, followed by help-steps to assist the child in accomplishing difficult items. A week later, the standard Cognition scale is administered again. The research was descriptive in nature and consisted of test administrations using the dynamic version with 57 children with developmental disabilities, along with interviews with six developmental psychologists. The results show that there is clear variability in the responses of children to the help-steps as well as in the score difference between the first and second administration of the standard Cognition scale. The developmental psychologists reported that the dynamic version clearly has added value for specific groups of children, such as children in the preliminary stage of school placement. The Task behavior questionnaire, which was part of the dynamic procedure, was evaluated positively by most educational psychologists. The conclusion is that the dynamic version of the Bayley-III-NL Cognition scale has added value for the developmental assessment of children with developmental disabilities.

This thesis concludes with **Chapter 7**, in which we discuss the results of the different studies and thereby answer the research questions. The limitations of the studies, along with the implications for early intervention practice are discussed, and we provide recommendations for future research. The research that was part of the current thesis has served to improve the possibilities for developmental assessment of young children with motor and/or visual, or language impairment(s). In addition, it has yielded an instrument that can provide additional useful information for setting up an intervention plan: the dynamic version. Together, the

Low Motor/Vision version, the Low Verbal version, and the dynamic version will be called the “Special Needs Addition” (SNA) and will be published in 2014 in the Netherlands as an addition to the standard version of the Bayley-III-NL. In spite of the fact that this will increase the number and quality of available instruments for developmental assessment, research remains necessary in order to continuously improve the existing instruments, as well as to develop instruments for other target groups. For children and adults with Profound and Multiple Learning Disabilities (PMLD), children with deafness, and children with profound motor impairment in the hands and arms, for example, the SNA does not contain sufficient accommodations, and a suitable instrument is thus still not available.

Samenvatting (Summary in Dutch)

De meeste kinderen die worden geboren, ontwikkelen zich volgens verwachting in de eerste jaren van hun leven. Bij een klein deel van de kinderen is echter iets aan de hand. Er kan sprake zijn van een lichte ontwikkelingsachterstand die het kind op een oudere leeftijd wel of niet inhaalt, maar ook van ernstigere ontwikkelingsproblemen. Voorbeelden van ernstigere problemen zijn aandoeningen die het gevolg kunnen zijn van problemen tijdens de zwangerschap of een te vroege geboorte (zoals Cerebrale Parese) of een genetische afwijking (zoals het syndroom van Down). Soms is de oorzaak onduidelijk, zoals vaak het geval is bij een algehele ontwikkelingsachterstand of een stoornis in het autistisch spectrum.

Als er sprake is van ontwikkelingsproblemen, is het belangrijk dat men daar achter komt als het kind nog jong is. Het blijkt dat behandeling op jonge leeftijd de meeste resultaten oplevert. Dit komt doordat kinderen zich erg snel ontwikkelen in de eerste jaren van hun leven. Door in die fase de ontwikkeling te stimuleren, kunnen de gevolgen van een beperking zoveel mogelijk beperkt worden. Het doel van een dergelijke interventie is een optimale ontwikkeling van het kind. Wat een optimale ontwikkeling is, is voor elk kind anders en hangt onder andere van de mate van beperking af. Als er bij ouders of hulpverleners het vermoeden bestaat dat er iets aan de hand is met een kind, vindt onderzoek plaats om erachter te komen wat er mogelijk aan de hand is. Een belangrijk onderdeel van dit onderzoek is ontwikkelingsdiagnostiek, waarbij wordt gekeken op welk niveau een kind zich ontwikkelt ten opzicht van leeftijdgenoten. Er wordt dan vaak naar verschillende gebieden van de ontwikkeling gekeken, zoals de verstandelijke ontwikkeling (cognitie), taal, motoriek en sociaal-emotionele ontwikkeling.

Instrumenten voor ontwikkelingsdiagnostiek worden ontwikkeld op basis van onderzoek met kinderen die een gemiddelde ontwikkeling doormaken, zodat ontwikkelingsnormen zijn op te stellen. Normen geven informatie over de verdeling van de testscores voor bepaalde leeftijdsgroepen. Vaak wordt echter niet nagegaan of het instrument ook de ontwikkeling van kinderen met een beperking op een goede manier kan onderzoeken. Dit blijkt niet altijd het geval. Op een goede manier onderzoeken houdt in dat het onderzoek betrouwbaar (nauwkeurig) moet zijn, maar ook valide, wat betekent dat er ook echt wordt onderzocht wat men wil onderzoeken. Als je bijvoorbeeld de cognitieve ontwikkeling onderzoekt, wil je niet meten of een kind bijvoorbeeld de motorische vaardigheden heeft om materiaal op te pakken. Je wilt meten of een kind het begrip heeft om bepaalde items (onderdelen van een test) uit te voeren. Er wordt echter vaak gebruik gemaakt van puzzeltjes of ander speelgoed dat niet geschikt is voor een kind met een motorische beperking. Als een beperking een kind belemmert een taakje goed uit te voeren, kan het kind niet laten zien of het begrijpt hoe het taakje moet worden uitgevoerd. Een onderschatting van het cognitieve ontwikkelingsniveau en vervolgens inadequate ondersteuning kunnen het gevolg zijn. Om dit te voorkomen, is het belangrijk om te zorgen dat de testmaterialen ook geschikt zijn voor kinderen met een beperking.

Een andere tekortkoming van veel instrumenten voor ontwikkelingsonderzoek is dat de testresultaten niet voldoende informatie geven die een basis kan vormen voor het ontwikkelen van een begeleidingsplan. Er is bijvoorbeeld informatie nodig over de begeleidingsbehoeften van een kind om tijdens een interventie goed te kunnen aansluiten bij het niveau en de behoeftes van het kind. Bestaande ontwikkelingstests blijken deze informatie niet in voldoende mate op te leveren.

In dit proefschrift wordt onderzoek beschreven dat tot doel heeft een instrument te ontwikkelen dat beter geschikt is voor ontwikkelingsonderzoek bij jonge kinderen met een beperking dan bestaande instrumenten zijn. De basis voor dit onderzoek is een veelgebruikt instrument voor ontwikkelingsonderzoek bij jonge kinderen: de Bayley-III-NL. Deze is ontwikkeld voor kinderen tot 3½ jaar oud en bestaat uit de schalen Cognitie, Taalbegrip, Taalproductie, Fijne Motoriek en Grove Motoriek.

Hoofdstuk 1 van dit proefschrift vormt de introductie, met een inleiding tot het onderwerp en een bespreking van de relevantie van het onderzoek voor de hulpverleningspraktijk. De onderzoeksvragen worden besproken en er wordt een overzicht gegeven van het gehele proefschrift.

Hoofdstuk 2 beschrijft een literatuuronderzoek naar beschikbare instrumenten voor ontwikkelingsonderzoek bij kinderen onder de 4 jaar. Er worden problemen beschreven die bestaan met betrekking tot ontwikkelingsonderzoek bij jonge kinderen met een beperking. Vervolgens wordt een overzicht gegeven van 18 instrumenten voor ontwikkelingsonderzoek die werden gevonden met behulp van het literatuuronderzoek. De geschiktheid van die instrumenten voor kinderen met een beperking wordt besproken. De conclusie is dat het aanbod aan instrumenten voor ontwikkelingsdiagnostiek beperkt is, vooral voor kinderen onder de 2 jaar.

Hoofdstuk 3 gaat over een pilotstudie naar de Low Motor/Vision versie van de Bayley-III-NL, die we hebben ontwikkeld om de Bayley-III-NL meer geschikt te maken voor kinderen met motorische en/of visuele beperkingen. Daarbij hebben we gebruikgemaakt van accommodaties: veranderingen aan het instrument die de invloed van de beperking op de testresultaten doen verminderen, maar die geen invloed hebben op datgene wat het instrument geacht wordt te meten. In het pilotonderzoek zijn 19 kinderen met een motorische en/of visuele beperking getest met zowel de Low Motor/Vision versie als de standaardversie. Uit de testresultaten blijkt dat een deel van de kinderen baat had bij de accommodaties. De testleiders gaven in een evaluatieformulier aan dat zij de accommodaties uitvoerbaar vonden en gunstig voor een meerderheid van de kinderen. De conclusie is dat de Bayley-III-NL Low Motor/Vision versie de ontwikkeling van kinderen met een motorische en/of visuele beperking op een valide manier lijkt te kunnen onderzoeken, maar dat uitgebreider onderzoek nodig is.

Hoofdstuk 4 beschrijft een uitgebreid onderzoek naar de Low Motor/Vision versie. De steekproef bestond uit 41 kinderen zonder beperking en 63 kinderen met een motorische en/of visuele beperking. Alle kinderen zijn met zowel de standaardversie als de Low Motor/Vision versie getest, in afwisselende volgorde. Uit de resultaten blijkt dat de kinderen zonder beperking vergelijkbare scores behaalden op de Low Motor/Vision versie en de standaardversie van de Bayley-III-NL. Dit ondersteunt het idee dat de accommodaties de inhoud en moeilijkheid van de items van de Bayley-III-NL niet hebben veranderd, zoals de bedoeling was. Op de Cognitieschaal behaalden de kinderen met een beperking gemiddeld een hogere score op de Low Motor/Vision versie dan op de standaardversie. Dit laat zien dat de accommodaties ervoor zorgden dat de kinderen hun vaardigheden beter konden laten zien. Het verschil in testscore tussen de twee versies werd niet gevonden voor de Taal- en Motoriekschalen. De testleiders gaven in een evaluatieformulier aan dat bijna alle kinderen hun vaardigheden goed konden laten zien met behulp van de Bayley-III-NL Low Motor/Vision versie en dat een meerderheid van de kinderen baat had bij de accommodaties. Voor een klein aantal kinderen was de Bayley-III-NL Low Motor/Vision versie niet geschikt, omdat de beperking te ernstig was en het kind nog steeds zijn/haar vaardigheden niet goed kon laten zien.

In **Hoofdstuk 5** beschrijven we een onderzoek naar de Low Verbal versie van de Bayley-III-NL. Deze versie bevat accommodaties voor de Cognitieschaal en de doelgroep is jonge kinderen met een spraak-/ taalbeperking. De Taal- en Motoriekschalen van de Low Verbal versie zijn niet aangepast. In het onderzoek zijn 69 kinderen met een spraak-/taalbeperking getest met de Bayley-III-NL Low Verbal en we hebben de resultaten vergeleken met die van een grote groep kinderen zonder beperking op de standaardversie. Hiervoor hebben we gebruik gemaakt van nonparametrische Item Respons Analyse (IRT). IRT is een statistische methode waarmee de eigenschappen van items van een test kunnen worden onderzocht. Uit de resultaten blijkt dat de items van de Bayley-III-NL Low Verbal, afgenomen bij kinderen met een spraak-/taalbeperking, dezelfde eigenschappen vertonen als die van de standaardversie, afgenomen bij kinderen zonder beperking. Uit de resultaten van een evaluatieformulier en interviews met ontwikkelingspsychologen uit de praktijk blijkt dat de Bayley-III-NL Low Verbal geschikt is voor kinderen met een spraak-/taalbeperking en bovendien geschikter dan de standaardversie in een meerderheid van de gevallen. De ontwikkelingspsychologen gaven aan de Low Verbal versie ook geschikt te vinden voor kinderen met een algehele ontwikkelingsachterstand. De conclusie is dat de Bayley-III-NL Low Verbal de ontwikkeling van kinderen met een spraak-/taalbeperking op een valide manier in kaart kan brengen.

Hoofdstuk 6 gaat over de dynamische versie van de Bayley-III-NL. We hebben een dynamische procedure toegevoegd aan de Cognitieschaal met als doel dat die relevante informatie oplevert voor een interventieplan. Na de standaardafname van de Cognitieschaal wordt er in bepaalde stapjes hulp gegeven bij taakjes die nog te moeilijk bleken. Een week later wordt de Cognitieschaal nogmaals afgenomen. Het onderzoek was beschrijvend en bestond uit testafnames met 57 kinderen met ontwikkelingsproblemen en interviews met zes ontwikkelingspsychologen. Uit de resultaten blijkt een duidelijke variabiliteit in de mate waarin de kinderen profiteerden van de verschillende hulpstappen en in het verschil in testscore tussen de eerste en tweede testafname. De ontwikkelingspsychologen gaven aan dat de dynamische versie duidelijk toegevoegde waarde heeft voor bepaalde doelgroepen, zoals kinderen in de fase van schoolplaatsing. De vragenlijst Taakgedrag, die onderdeel is van de dynamische versie, werd door de meeste ontwikkelingspsychologen positief geëvalueerd. De conclusie is dat de dynamische versie van de Bayley-III-NL Cognitieschaal van toegevoegde waarde is voor ontwikkelingsonderzoek bij kinderen met ontwikkelingsproblemen.

Dit proefschrift sluit af met **Hoofdstuk 7**, waarin de resultaten van de verschillende onderzoeken worden samengevat en daarmee een antwoord wordt gegeven op de onderzoeksvragen. De beperkingen van het onderzoek en de implicaties voor de praktijk worden besproken en er wordt een aanzet gegeven voor toekomstig onderzoek. Het onderzoek heeft de mogelijkheden voor ontwikkelingsonderzoek verbeterd voor jonge kinderen met motorische en/of visuele, of spraak-/taalbeperking(en). Daarnaast heeft het een instrument opgeleverd dat aanvullende nuttige informatie kan geven voor het ontwikkelen van een interventieplan: de dynamische versie. De Low Motor/Vision, Low Verbal en dynamische versie zullen gezamenlijk de “Special Needs Addition” (SNA) worden genoemd en met die naam in 2014 in Nederland worden uitgegeven als bijlage bij de standaardversie van de Bayley-III-NL. Ondanks het feit dat dit het aanbod van geschikte instrumenten voor ontwikkelingsonderzoek bij kinderen met een beperking verbetert, blijft onderzoek hard nodig. Dat zal zich moeten richten op het continu verbeteren van de bestaande instrumenten voor ontwikkelingsdiagnostiek, maar vooral ook op het ontwikkelen van geschikte instrumenten voor andere doelgroepen, zoals kinderen en volwassenen met Ernstige Meervoudige Beperkingen (EMB), dove kinderen, blinde kinderen en kinderen met zeer ernstige motorische beperkingen in de handen en armen. Voor deze doelgroepen zijn de accommodaties in de SNA niet voldoende.

Dankwoord (Acknowledgements)

Het vierjarige “Bayley SNA-project” zit erop. Het resultaat van dat project is niet alleen dit proefschrift, maar ook een kast met ruim 500 dossiers waarvan de inhoud in de praktijk gebruikt is. Het meest belangrijke product is de aangepaste versie van de Bayley-III-NL en de kennis die we daarover hebben opgedaan. Deze komt de hulpverleningspraktijk en daarmee de begeleidingsmogelijkheden voor jonge kinderen met ontwikkelingsproblemen ten goede.

Eén van de vele dingen die ik in de afgelopen jaren heb geleerd, is dat je een proefschrift echt niet kunt schrijven zonder de hulp van anderen. In dit stuk wil ik de mensen bedanken die een belangrijke rol hebben gespeeld in het Bayley SNA-project en bij de totstandkoming van dit proefschrift. Behalve een dankwoord, vormt dit daarmee ook een overzicht van de mensen die betrokken zijn geweest en de rol die zij gespeeld hebben.

Om te beginnen wil ik ZonMw bedanken, die het hele project gefinancierd heeft. Ook Pearson speelt een belangrijke rol, omdat de uitgave van de SNA-versie ervoor zorgt dat het instrument echt structureel in de praktijk gebruikt kan gaan worden. Daarnaast hebben de deelnemende instellingen, de betrokken orthopedagogen en natuurlijk de ouders en kinderen die hebben deelgenomen een onmisbare bijdrage geleverd. Zonder deze deelname was het project niet mogelijk geweest. Bedankt allemaal!

Het “Bayley SNA-projectteam”, tegelijk mijn begeleidingsteam, werd gevormd door Selma, Bieuwe, Marieke en Wied. Ik wil jullie heel erg bedanken voor de fijne samenwerking! Zowel voor het project als voor mijn begeleiding vulden jullie elkaar heel goed aan en daardoor is het geheel niet alleen tot een goed einde gekomen, maar heb ik ook veel geleerd. Het was heel fijn dat jullie altijd beschikbaar waren als dat nodig was en mij, naast een goede sturing, door regelmatige complimenten veel vertrouwen hebben gegeven. Dit heeft mij goed gedaan en geeft niet alleen vertrouwen in de kwaliteit van het nu afgeronde project, maar ook voor de toekomst.

Selma, een groot deel van de dataverzameling hebben we samen gecoördineerd. In het begin jij een beetje meer dan ik, later andersom. Ik heb echt

veel van je geleerd, onder andere over het uitvoeren van wetenschappelijk onderzoek in de praktijk. Ik denk dat het in het SNA-project veel heeft geholpen dat je het belang van de praktijk en de doelgroep nooit uit het oog verliest. Dat komt terug in de manier waarop je de mensen in de praktijk benadert, altijd zoekend naar manieren om het voor hen zo aantrekkelijk mogelijk te maken om deel te nemen.

Bieuwe, ik heb niet vaak iemand gezien die het zo druk heeft terwijl hij al met 'pensioen' is! Heel fijn dat het SNA-project één van de dingen was waar je zo druk mee was. Bij het schrijven van de teksten hield je altijd de theoretische basis in het oog en zorgde je er bijvoorbeeld voor dat ik de begrippen goed definieerde. Het gezamenlijke bezoek aan het congres in Sint Petersburg zal ik niet snel vergeten.

Marieke, behalve je bijdrage aan het statistiekdeel van het onderzoek, was je erg betrokken en je heldere kijk op de zaken was bij de vele projectbesprekingen erg nuttig. Ik heb veel van je geleerd door je duidelijke uitleg, het gezamenlijk papers schrijvers en je feedback op teksten. Ik vind dat we op een hele fijne manier hebben samengewerkt. Misschien dat dat komt doordat we wel wat op elkaar lijken wat betreft (een gestructureerde) manier van werken.

Wied, hoewel je begeleiding wat meer op afstand was, bleek dat juist van grote waarde, doordat je met een iets andere blik naar het project kon kijken. Het was fijn altijd bij je te kunnen binnenwandelen, of het nu ging om een handtekening of een vraag om advies over bijvoorbeeld cursussen. Ook was het fijn altijd een hoopgevend dan wel vrolijk mailtje te ontvangen nadat een tijdschrift de uitslag van een review bekend had gemaakt ("Ik ben weliswaar op vakantie, zit boven in de Alpen en lees geen mails, maar dit kan ik niet laten gaan natuurlijk: gefeliciteerd!").

We hebben van heel wat mensen hulp gehad tijdens het onderzoek. Yvonne was als eerste onderzoeksstagiaire betrokken en heeft vervolgens als onderzoekmedewerker vele testafnames gedaan, de dossierkast gesystematiseerd en die een lange tijd bijgehouden. Bedankt voor je vele hulp! Je vriend, Ulbe Cees, heeft de beroemde "Bayley beheertool" ontworpen, waarvan we tijdens de tweede helft van het onderzoek erg veel baat hebben gehad. Bineke heeft in de tweede helft van het project vrijwillig meegeholpen met onder andere testafnames, kinderen werven voor de controlegroep en testgegevens invoeren in SPSS. Heleen, Elvera en Kitty als onderzoeksstagiaires en Ineke, Juliet, Elsbeth, Jeannette, Evelien, Marianne en Sigrid als thesisstudenten hebben veel bijgedragen aan de ontwikkeling van het testmateriaal en het afnemen van de tests voor het onderzoek. Daarnaast waren de meesten van hen ook testleider voor en/of na hun masterproject. Ook heeft een heel

aantal studenten in het kader van hun bachelorproject geholpen bij het ontwikkelen van de materialen en testen van de kinderen. Allemaal heel hartelijk bedankt!

Dan de testleiders: Dewi, Doety, Elizabeth, Els, Grytjse, Hannah Beth, Hanneke, Inez, Leaniek, Loes, Lotte, Maaike, Marijn, Marinda, Marleen, Marlies, Martine, Michell, Nica, Romy, Sharon, Simone, Susan, Thijs en Wilma. Jullie hebben je vrijwillig ingezet voor het testen en zonder jullie was het project niet uitvoerbaar geweest. Ook jullie heel erg bedankt!

Ik heb het in de afgelopen jaren erg naar mijn zin gehad op de afdeling orthopedagogiek. Ik wil graag alle collega's bedanken, die samen de afdeling een leuke plek om te werken maken. Specifiek wil ik de Aio's / jonge medewerkers bedanken. Met name in de eerste periode, maar ook daarna, was het heel fijn om altijd bij collega's terecht te kunnen met vragen en niet alles zelf uit te hoeven zoeken. Daarnaast waren de Aio-uitjes en vooral ook de gezamenlijke pauzes (die vooral Tim er goed in houdt!) heel gezellig en een goede stok achter de deur om af en toe even pauze te nemen. Meenakshi and Suzanna, thanks for the nice Monday evenings we have spent (and hopefully will spend) together! Nynke, Els, Kim, Ineke en Josien, bedankt voor de gezellige momenten, ook buiten werktijd! Nynke ook bedankt voor de leuke tijd samen op de kamer. Het was fijn dat we hard werken goed konden afwisselen met gezelligheid. En tenslotte nog: Nynke en Ineke, leuk dat jullie met zo veel enthousiasme mijn paranimfen willen zijn!

In het kader van de promotie zelf wil ik de leescommissie bedanken voor de bereidheid om deel te nemen aan de commissie.

Aan het 'thuisfront' ga ik niet veel woorden wijden, omdat de mensen die voor mij belangrijk zijn, dat wel weten. Maar om het overzicht van mensen met een belangrijke rol in mijn promotietraject compleet te maken: Mathieu, je hebt mij niet alleen gesteund en gestimuleerd om deze promotie te doen, ondanks het feit dat je je er in bepaalde opzichten behoorlijk voor hebt moeten aanpassen. Je was ook altijd degene die het als eerste merkte als het druk was of ik het overzicht even niet meer had. En je was er dan om me erop te wijzen dat ik daar misschien iets aan kon doen... Dank je wel!

Curriculum Vitae

Linda Visser was born on May 21, 1985, in Woerden (the Netherlands). She attended primary and secondary school in Woerden. In 2003, she started her Psychology studies at Utrecht University. During her studies, she specialized in the area of child and adolescent psychology and conducted her placement at child day care centre *'t Steyntje* for young children with developmental disabilities. After finishing her Master degree in 2007, she moved to Dublin (Ireland), where she worked as a care assistant in respite care for children with disabilities. From January 2010 until January 2014 she worked on her PhD project. Currently she is working as assistant professor at the University of Groningen.